Research Product 89-20

Questionnaire Construction Manual

June 1989

Fort Hood Field Unit
Systems Research Laboratory

U.S. Army Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

JON W. BLADES
COL, IN
Commanding

Research accomplished under contract for the Department of the Army

Essex Corporation

Technical review by

David Meister
W.F. Moroney, MSC, U.S. Navy

NOTICES

FINAL DISPOSITION: This Research Product may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: This Research Product is not to be construed as an official Department of the Army document, unless so designated by other authorized documents.

REPRODUCTION QUALITY NOTICE

This document is the best quality available. The copy furnished to DTIC contained pages that may have the following quality problems:

- Pages smaller or larger than normal.
- Pages with background color or light colored printing.
- Pages with small type or poor printing; and or
- Pages with continuous tone material or color photographs.

Due to various output media available these conditions may or may not cause poor legibility in the microfiche or hardcopy output you receive.

☐ If this block is checked, the copy furnished to DTIC contained pages with color printing, that when reproduced in Black and White, may change detail of the original copy.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS --
2a. SECURITY CLASSIFICATION AUTHORITY --	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE --	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) --	5. MONITORING ORGANIZATION REPORT NUMBER(S) ARI Research Product 89-20
6a. NAME OF PERFORMING ORGANIZATION Essex Corporation Human Factors & Training Systems Group 6b. OFFICE SYMBOL (If applicable) --	7a. NAME OF MONITORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences
6c. ADDRESS (City, State, and ZIP Code) 741 Lakefield Road, Suite B Westlake Village, CA 91361	7b. ADDRESS (City, State, and ZIP Code) ARI Field Unit at Fort Hood HQ TCATA (PERI-SH) Fort Hood, TX 76544
8a. NAME OF FUNDING/SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences 8b. OFFICE SYMBOL (If applicable) PERI-ZA	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER MDA903-83-C-0033
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: 63739A PROJECT NO.: 793 TASK NO.: 321 WORK UNIT ACCESSION NO: 0

11. TITLE (Include Security Classification)

Questionnaire Construction Manual

12. PERSONAL AUTHOR(S)

Babbitt, Bettina A. (Essex Corporation), and Nystrom, Charles O. (ARI)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 84/12 TO 85/03	1989, June	227

16. SUPPLEMENTARY NOTATION (Continued)

This is a revised version of the July 1976 Questionnaire Construction Manual, P-77-1, originally authored by R. F. Dyer, J. J. Mathews, C. E. Wright, and K. L. Yudowitch.

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

Questionnaire construction, Scaling techniques, Questionnaire administration, Response anchoring, Attitude scales. (Continued)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This Questionnaire Construction Manual is a revised version of a 1976 manual. The latest research methods for developing questionnaires are presented. The manual was designed to guide individuals who develop and/or administer questionnaires as part of Army field tests and evaluations. The content is applicable to many nonmilitary applications.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☑ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Charles O. Nystrom	22b. TELEPHONE (Include Area Code) (817) 288-9118 22c. OFFICE SYMBOL PERI-SH

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

ARI Research Product 89-20

16. SUPPLEMENTARY NOTATION (Continued)

Operations Research Associates, Palo Alto, CA. Charles O. Nystrom is the Contracting Officer's Representative.

18. SUBJECT TERMS (Continued)

Response alternatives
Pretesting questionnaires
survey interviews

Research Product 89-20

Questionnaire Construction Manual

Bettina A. Babbitt
Essex Corporation

and

Charles O. Nystrom
U.S. Army Research Institute

Field Unit at Fort Hood, Texas
George M. Gividen, Chief

Systems Research Laboratory
Robin L. Keesee, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

June 1989

Army Project Number
2Q263739A793

Human Factors Evaluation

FOREWORD

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), Field Unit at Fort Hood, Texas, actively guided this revision of their 10-year-old Questionnaire Construction Manual (P-77-1). The questionnaire construction manual was designed to guide individuals who develop and/or administer questionnaires as part of Army operational tests. It is, however, suitable for a variety of disciplines and occupations. Guidance is provided in the development of questionnaire items, administration procedures, types of questionnaire items, attitude scales and scaling techniques, response anchoring and response alternatives, format considerations, pretests, interviews, demographic characteristics, and evaluation of results.

This product was completed under Program Task 1.5.1, "Soldier/System Considerations in Force Development User Testing (Advanced Development)." ARI and the Sponsor for the product work under a "Memorandum of Agreement between ARI and Training and Doctrine Command (TRADOC) Combined Arms Test Activity (TCATA)" that was signed in May 1981. The Chief of TCATA's Methodology and Analysis Section has been briefed on the product content. TCATA has been using the predecessor Questionnaire Construction Manual to test officers for over 10 years and would like to use the updated product.

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

Several people helped to prepare this manual. A special acknowledgment goes to Dr. Frederick A. Muckler, Essex Corporation, for his guidance and continuous support during all aspects of the preparation of this report. The contribution of Mr. George M. Gividen, U.S. Army Research Institute for the Behavioral and Social Sciences, is most gratefully acknowledged. Mr. Clarence A. Semple, Essex Corporation, contributed generously in editing. Mrs. Joan M. Funk, Essex Corporation, merits special recognition for her technical assistance in preparing and editing the manuscript.

EXECUTIVE SUMMARY

This manual updates the 10-year-old "Questionnaire Construction Manual." The revision was prepared primarily by the Essex Corporation under contract to the Army Research Institute for the Behavioral and Social Sciences (ARI). It has the same purpose as the earlier version--to provide guidance for those who construct and/or administer questionnaires as part of Army operational tests and evaluations such as those conducted by the TRADOC Combined arms Test Activity and the Operational Test and Evaluation Agency. Much of the content is applicable to more than operational test situations; the manual should prove useful to all persons involved in the construction and administration of surveys, interviews, or questionnaires.

In 1975, Operations Research Associates revieved the research literature on the construction and administration of questionnaires and interviews. They produced two products. One was the forerunner of this manual. It was titled "Questionnaire Construction Manual" and was published by ARI in 1976. A revision was done in 1976 and issued in quantity in 1977 as ARI Special Publication P-77-1. The other product was a report of the literature survey and a bibliography of the articles examined. It was issued in 1977 as P-77-2, with the title "Questionnaire Construction Manual Annex: Literature Survey and Bibliography."

In 1983, the literature was again revieved, but only from the point where ORA's review had ended in 1975. Analysis of the more recent literature provided the basis for the revision to the manual. A report of the literature survey has been published under the title, "Questionnaires: Literature Survey and Bibliography."

QUESTIONNAIRE CONSTRUCTION MANUAL

CONTENTS

CONTENTS (Continued)

CONTENTS (Continued)

LIST OF TABLES

LIST OF FIGURES

QUESTIONNAIRE CONSTRUCTION MANUAL

Chapter I: Introduction

A. Purpose and Organization of This Manual

1. Purpose

This manual has been prepared primarily for the use and guidance of those who are tasked to develop and/or administer questionnaires as part of Army field tests and evaluations, such as those conducted by the TRADOC Combined Arms Test Activity (TCATA), the Combat Developments Experimentation Command (CDEC), the Operational Test and Evaluation Agency (OTEA), and the several Army Boards and Schools. The general content and concepts, however, are applicable to a variety of situations. As such, the manual should prove useful to all individuals involved in the construction and administration of surveys, interviews or questionnaires.

2. Organization

Information and guidance relating to the preparation of items for questionnaires and for their assembly and arrangement into a complete questionnaire are presented in Chapters II through X. Chapter XI discusses the importance of, and procedures for, pretesting questionnaires prior to their regular administration. Chapter XII discusses characteristics of respondents that influence questionnaire results. The analysis and evaluation of responses to a questionnaire are briefly dealt with in Chapter XIII. Finally, a number of considerations regarding the presentation of questions by means of an interview are discussed in Chapter XIV.

B. Definition of Questionnaire

As used in this manual, the word "questionnaire" refers to an ordered arrangement of items (questions, in effect) intended to elicit the evaluations, judgments, comparisons, attitudes, beliefs, or opinions of personnel. The content and format of the items may vary widely. A visual mode of presenting the items is employed. In the past, this meant that the items were typed or printed on paper, but now items can also be presented by closed circuit television or on a cathode ray tube (CRT) or on a video display terminal (VDT) under the control of a computer program. If the items are first read by an interviewer and then given verbally to the respondent, the questionnaire may also be termed a "structured interview." Hence, questionnaires and interviews have some common properties. Questionnaire items used to be responded to by scribing words or marks with a pen or pencil, but this aspect too has been enlarged to include typed, punched, button-pushing, light-penned, joystick, and verbal responses.

While questionnaires are "data collection forms," not all data collection forms are questionnaires. Those forms used by personnel to enter instrument readings or to record their counts or observations (e.g., time of first detection, number of targets correctly identified, number of rounds fired) are not directly addressed in this manual.

C. Conventions Used in This Manual

1. Identification Scheme Used

This manual has been prepared in outline form to facilitate cross-referencing and later updating. The identification scheme that is used employs Roman numerals, capital and small letters, and numbers in the sequence: I A 1 a (1) (a) [1] [a]. The major divisions, I, II, III, IV, etc., are called chapters. All other subdivisions are called "sections," with sections starting with capital letters (A, B, etc.) called "major sections." You are now, for example, reading Section I-C 1. To facilitate later updating, references within the manual are to sections and not pages.

2. Pagination

Each major section of this manual (e.g., I-C) starts on a new page, and pages are numbered within each major section. For example, this is Section I-C Page 1, or the first page of Section I-C.

3. Page Update Date

Immediately under each page number is the date that the page was drafted or revised. When a page has been revised, the date of the immediately previous version is also given in parentheses with the letter "s" meaning superseded." For example, III-B Page 1 dated 1 Jul 76 was revised on 8 Mar 85. The page number on the revised page would appear as:

III-B Page 1
8 Mar 85
(s. 1 Jul 76)

When updating the manual, new material that was not previously part of the text would not require the letter "s," For example, IV-E Page 6 originated on 8 Mar 85 would appear as:

IV-E Page 6
8 Mar 85

4. Table and Figure Identification

Both tables and figures are numbered sequentially within a major section, with a hyphen before the table or figure number. Examples are: Table VIII-B-1, Table VIII-B-2, Figure VI-A-1.

D. Keeping This Manual Up to Date

1. Updated Pages Should be Inserted as Received

It is anticipated that sections of this manual will be periodically corrected, revised, or otherwise updated. New pages should be inserted as soon as they are received. This will not only keep the manual up to date, but will facilitate adding pages received at an even later date. Appropriate instructions covering which pages to add and delete will accompany distributed update pages. When it appears useful, a list will also be provided showing the page numbers and dates of all pages that should be in the manual at that time.

2. Request for Updates

To be placed on the distribution list to receive updates to this manual, write to:

Chief
ARI Field Unit-Fort Hood
HQ TCATA (PERI-OH)
Fort Hood, Texas 76544-5065

E. Reporting Problems and Suggestions for Improvement

As previously noted, it is anticipated that this manual will periodically be updated to improve its utility. To report errors, problems, or suggestions, write to:

Chief
ARI Field Unit-Fort Hood
HQ TCATA (PERI-OH)
Fort Hood, Texas 76544-5065

Chapter II: Major Questionnaire Types and Administration Procedures

A. Overview

This chapter briefly summarizes the different types of questionnaires discussed in this manual (Section II-B) and ways that questionnaires may be administered (Section II-C). Detailed guidelines regarding what to do in a given situation are included in subsequent chapters. Issues to consider when deciding whether to use a structured interview or some other type of questionnaire are presented in Section II-D, which also notes that combinations of methods may be employed. It is concluded that both structured interviews and other types of questionnaires have their place. Each has strengths and limitations which must be taken into account when identifying which instruments to use.

B. Types of Questionnaires Discussed in This Manual

There are a number of techniques of data collection that can be used to measure human attributes, attitudes, opinions, and behavior. Attitude and opinion are closely aligned if not overlapping. Opinions are restricted to verbalized attitudes. Attitudes are sometimes unconscious or nonverbalized. Some of the methods of data collection are observation, personal and public records, specific performances, sociometry, interviews, questionnaires, rating scales, pictorial techniques, projective techniques, achievement testing, and psychological testing. For this manual, however, attention has been restricted to a more limited number of data collection techniques: certain paper-and-pencil types of instruments broadly classed as questionnaires as defined in Section I-A 2, and including only some of the techniques mentioned above. A distinction has also been made in this manual between open-ended questionnaire items and closed-end items. Open-ended items are those which permit respondents to express their opinions in their own words, and to indicate any qualifications they wish. The amount of freedom the respondent will be given in expressing an answer to an open-ended item is partly determined by the questionnaire designer. Closed-end items use response alternatives. Respondents are directed to select one or more of the response alternatives from a closed set. Closed-end items frequently used are multiple choice, true-false, checklist, rating scale, and forced-choice. Survey items have been roughly classified into two groups: open-ended items and closed-end items.

It is common to use interview surveys to ask questions and record answers. Structured interviews are included within the definition of questionnaires used, since typically an interview form is developed and used by an interviewer both for asking questions and recording responses, much like a self-administered questionnaire. On the other hand, the unstructured interview makes no use of structured data collection forms. The interviewers are permitted to discuss the subject matter as they see fit with no particular order or sequence. Of course, other interviews fall somewhere between these two extremes. In any case, unstructured interviews, where no structured response forms are used, are <u>not</u> included within the definition of questionnaires used in this manual.

C. Ways That Questionnaires Can Be Administered

There are a number of respects in which questionnaire administrations may vary. However, in the usual field test settings, the typical questionnaire administration situation involves paper-and-pencil materials with the author/test officer administering the questionnaire face-to-face with a group of test players or evaluators.

1. Group Versus Individual Administration

Given a printed questionnaire, calendar time is saved by group administration. Group administration allows the opportunity for a questionnaire administrator to explain the survey and answer questions about items. The task of statistical analysis can be initiated with less delay than if one were waiting on a series of individual administrations. An important determinant of group vs. individual is the time at which people complete their participation in the test. Most often all participants are through at the same time. All would be available for questionnaire administration as soon as they could be brought to an appropriate place or places. Prompt group administration gives the same short amount of time for forgetting about test events by those who become the respondents. Group administration generally has a high cooperation rate. If there is an administrator, his/her time is conserved directly in proportion to the number of respondents he/she has in each administrative session. An advantage of group administration is low cost.

2. Author-Administered Questionnaires

When the test officer or administrator who is familiar with the content of the questionnaire and the test's purposes/objectives can administer the questionnaire, some advantages can be gained. The administrator's instructions and appeals may increase the number of respondents having desirable motivation to complete the questionnaire by giving appropriate consideration to each item. If one employs a self-administration procedure, such as might occur in a mailed-out questionnaire, or if a poorly prepared stand-in plays the role of administrator, then the respondents must derive their instructions and some of their motivation from printed instructions (or from the poorly prepared stand-in). More things usually can end up going wrong when questionnaires are self-administered than when they are administered by a test administrator.

3. Remote Administrations

From the test officers' point of view, remote administration refers to a questionnaire administration event that they cannot conduct because of its distance from them and/or other demands on their time. This dimension, remote versus face-to-face, is similar but not identical to the previously noted dimension, self-administered versus author administered.

To avoid the possible disadvantages of self-administered questionnaires, the test officer must be able to afford another administrator, train him/her in the knowledge and skills associated with effective administration, and transport him/her to the "remote" administration location. If multiple administrations having location or timing differences which preclude the same administrator from handling them are required, it would appear that the chances are increased that more respondents will experience more "difficulties" in answering the questions. For this type of questionnaire administration, the questionnaire itself would require careful design associated with items and instructions.

4. Other Materiel Modes

Providing the respondents with a printed questionnaire form, and a pencil to mark/write their responses, is the most common questionnaire administration procedure in field evaluations. In addition, other presentation modes have been used. In a card-sorting procedure that has been used with individuals and groups, each respondent reads statements of candidate problems and then places the card into the appropriate pile according to his/her judgment of the severity of the "problem." Rarer because of expense and logistics problems is the setting up of a computer terminal where each respondent enters (types in) answers to questions that are displayed on a cathode ray tube (or other computer display device). Chapter XII presents many other considerations related to questionnaire administration.

D. Structured Interviews Versus Other Types of Questionnaires

1. Issues to Consider

When deciding whether to use a structured interview or another type of questionnaire, a number of issues should be considered.

Included are the following:

a. To develop questionnaire items, a focus group may be interviewed. Their comments can be used to develop hypotheses and refine questions. This information can be adapted to an interview guide and interview items.

b. Interview items should not use a dichotomous response set. Multiple choice and open-ended questions provide the opportunity for probing.

c. If a structured interview is used, there must be enough qualified interviewers to expeditiously process all interviewees. Sometimes there are only a few personnel to be interviewed, or there is plenty of time available for interviews, so only one or two interviewers will be necessary. In other situations, maybe only an hour or so may be available per interviewee; in these cases, a large number of qualified interviewers must be available.

d. Face-to-face interviews have a higher response rate than mail surveys.

e. In most cases, respondents have a greater tendency to answer open-ended questions in an interview than when response is by paper and pencil.

f. It is possible to adapt face-to-face interview guides for telephone surveys. Oral labeling of the scale points should be assessed on a pilot survey to be sure that the responses are not biased by the oral presentation of the scale.

g. Telephone interviews are faster to perform than mail surveys.

h. Interviews conducted by telephone require an interview structure that promotes a high interaction between the interviewer and respondent.

i. Group-administered paper-and-pencil questionnaires may be less expensive, more anonymous, and completed faster than the same number of interviews.

j. Respondents seem to be less likely to report unfavorable things in an interview than in an anonymous questionnaire. Typically, questionnaires are also more likely than interviews to produce self-revealing data.

k. Issues involving socially acceptable or unacceptable attitudes and behaviors will elicit more response bias.

l. During interviews, respondents often have a tendency to try to support the norms that they assume the interviewer adheres to.

m. Interviewers with biases on the issues under discussion may reflect them in the content they record, as well as in what they fail to record.

n. Ethnic background differences between interviewer and respondent probably will not influence the survey results unless the items have a racial content or are found to be threatening.

o. Although a structured interview using open-ended questions may produce more complete information than a typical questionnaire containing the same questions, empirical research seems to indicate that responses to the typical questionnaire are more reliable; i.e., more consistent. Structured interviews using closed-end questions appear to be as reliable as paper-and-pencil questionnaires.

p. It may be difficult to code a combination of open-ended and closed-end items for interview surveys. (See Section XIII-B, Scoring Questionnaire Responses.)

2. Combinations of Methods

There are some situations where a combination of methods of questioning might be used:

a. An interview might be used to obtain information for designing a paper-and-pencil questionnaire.

b. Personal interviews or telephone interviews might be used for respondents who do not return questionnaires administered remotely (such as mail questionnaires).

c. When respondents are unable to give complete information during an interview, they can be left a copy of a questionnaire to complete and mail in, so that the necessity for a call-back is eliminated.

3. Conclusion

Both structured interviews and other types of questionnaires appear to have their advantages and disadvantages. The choice of which to use may well depend upon costs, which are generally lower for the typical questionnaire. The typical questionnaire is apparently more reliable, while the structured interview may provide more unique and more abundant information. If the dimensions of a problem have not been explored before, the best compromise would appear to be to use the interview approach with open-ended items to uncover the dimensions, and follow this by the use of the paper-and-pencil questionnaire with closed-end items to obtain more specific information.

Chapter III: Content of Questionnaire Items

A. Overview

The recommended general steps in preparing a questionnaire include preliminary planning, determining the content of questionnaire items, selecting question forms, wording of questions, formulating the questionnaire, and pretesting. As part of preliminary planning, the information required has to be determined, as do procedures required for administration, sample size, location, frequency of administration, experimental design of the field test, and analyses to be used. Selecting question forms is a function of the content of the questionnaire items and requires knowledge of types of questionnaire items and scaling techniques. The wording of questions is the most critical and most difficult step. Formulating the questionnaire includes formatting, sequencing of questions, consideration of data reduction and analysis techniques, determining basic data needed, and insuring adequate coverage of required field test data. Pretesting involves using a small but representative group to insure that all questions are understandable and unambiguous.

This chapter considers the content of questionnaire items. Methods for determining questionnaire content are discussed first, and then other considerations related to questionnaire content are presented. The other steps noted above are discussed in subsequent chapters.

B. Determining Questionnaire Content Preliminary Research

1. Preliminary Research

If you have the job of developing a questionnaire for a field test, there are several things that should be done before starting to write questionnaire items.

a. Learn the test's objectives and issues. Read the Outline Test Plan in order to learn what it says the test's purpose, scope, and objectives are. All data collection effort, including questionnaire administration, should be consistent with and supportive of the test's objectives. Read the Independent Evaluation Plan, with its discussion of issues and of ways of collecting data on the issues.

b. What performance measures are planned for the test? One may be fortunate enough to be involved with a test for which the Detailed Test Plan has to a large extent been written. Try to discover what performance measures/data are to be collected. If performance data is to be collected on some aspects of the functioning of the system to be tested, then it may not be necessary to assess these functions via questionnaire items. Make a list of what should be measured to meet the objectives of the field test. The list will include variables that are configured into categories. The list should not include any questions.

c. Consult others and prior test plans and reports. Many tests at CDEC and TCATA (and elsewhere) follow-up, or are similar to, prior testing. As a consequence, information may be readily available regarding prior related or similar tests. Test files or the Technical Information Center may provide a source for obtaining test plans and reports on relevant prior tests conducted by Army field test/experimentation agencies.

d. Consult others and develop an analysis plan. The Technical Information Center may provide guidance for data analysis. Develop an analysis plan with a list of variables to be measured. The analysis plan identifies dependent and independent variables. It also identifies which variables to control and any intervening variables.

Preliminary research requires an understanding of the objectives of the test plan, a list of the variables to be measured, and a plan for analysis of the data.

2. Using Interviews to Determine Questionnaire Content

If one's degree of experience seems meager relative to the complexities of the evaluation problem, he/she may employ group and/or individual interviews to assist in determining questionnaire content. Preferably, this would be done after taking the steps noted above. The less one knows about a subject, the less structure one can impose on an interview dealing with the subject.

a. Conducting an unstructured group interview. Personnel are needed who have relevant operating experience with the system to be tested/evaluated - or with a sufficiently similar system. Arrange a common meeting place and time with about five to ten of them. It would be advantageous to have a meeting place that was not cramped for space, had comfortable chairs, a comfortable temperature, and where all discussants were free from other sources of distraction (sights and sounds, mainly).

If the interviewer's age and rank are several steps above or below the age and rank of the members of a homogeneous group of discussants, try (before the meeting) to get a person who is their contemporary (peer) in age and rank to lead and coordinate the discussions. Why? Because a mismatch may inhibit their discussion or produce too much submissive, agreeing behavior on their part.

If notes are being taken or the discussion is being tape recorded, one should be unobtrusive about it. Don't shove/point a microphone at people as they start to speak. They may be inhibited by this, or they may become "hams."

The first several minutes should be spent in establishing rapport with the group. The purpose of the session should be covered, introduction of group members made, and other warm-up devices used. The objective is to motivate as many respondents to give comments as possible. In the remainder of the session, any or all of the following information-eliciting devices could be used:

(1) Discuss samples of the control item--ask the general question: "What problems have you had with this piece of equipment or system?" Follow up with who, what, where, when and why. Attempt to maximize the number of potential or actual problems posed. Strive for clarification of problem ideas, but do not criticize the comments, even if they are redundant with a previous contribution by the respondent or other respondents.

(2) Ask: "What do you consider to be the most important features (characteristics, qualities, etc.) of this equipment or system when used in the field?" Strive to get a multitude of adjectives and phrases here (e.g., ease of operation, weight, durability, portability, etc.).

(3) Use the aided recall technique: "Can you remember where and when you have encountered problems with this system?" (e.g., at night; when it's damp, etc.).

(4) The way survey issues are discussed will help in selecting vocabulary and phrasing questions.

(5) Researchers interested in obtaining accurate data from their interviews generally ask multiple questions for each topic. The questions are sequenced to provide smooth transitions throughout the interview. Development of questionnaire items is based on hypotheses that have been developed. The hypotheses are presented to a group of individuals who are subject matter experts, and they perform a preliminary assessment of the hypotheses. The questionnaire may require modification if the hypotheses are not viable.

The recorded comments should be categorized and arranged by frequency. For example, how many of the comments on system operation stressed failure considerations?

b. Conduct semistructured personal interviews. Information produced from the unstructured group interviews provides general guidance to the specific evaluative information desired. As a next step, or as an alternative step to the group interview, one may employ a small number of representative respondents in a person-to-person interview format.

In this method of interviewing, the interviewers are given only general instructions on the type of information desired. They are left free to ask the necessary direct questions to obtain this information, using the wording and the order that seems most appropriate in the context of each interview. These interviews, like the unstructured group sessions, are useful in obtaining a clearer understanding of problems, and in determining what areas (evaluation criteria) should be included on the pilot questionnaire.

The only structure to the semistructured interview comes from a set of question categories that must be raised sometime during the interview. Questions on system experience, positive and negative features, and problems in field use, for example, can be phrased in any manner or sequence. Probing questions of the type: "Why do you feel that way?," "What do you mean by that statement?," and "What other reasons do you have?" can be utilized until the interviewers are satisfied that they have the necessary information considering time limitations, data requirements, and the willingness and ability of the respondents to verbalize their views. Interview forms should be designed to allow the interviewer sufficient space for writing notes and comments.

In the semistructured interview, the interviewer has some flexibility in formulating and asking questions. This technique can, therefore, be only as effective in obtaining complete, objective, and unbiased information as the interviewer is skilled in formulating and asking questions. Thus, interviewers may have to be trained in using this technique.

When interviews are used as the basis for a future questionnaire, the questions need to be carefully stated so that they are eliciting data which will enable the interviewer to construct questions which address the stated objectives and issues of the research. Once the questionnaire items have been identified, the items need to be assembled into a logical sequence. They then need to be administered to a sample of respondents who have a background similar to the audience to which the questionnaire was originally targeted. Information obtained from the sample administration is used to refine questionnaire items.

c. Develop the questionnaire. In the development phase of a questionnaire, an open-ended response format can be useful in selecting meaningful response alternatives for a multiple choice format. Open-ended questions administered to a sample of the target population will provide responses that can then be phrased in the spontaneous wording of the individuals in the sample. The questionnaire items can be pretested using an open-ended response format on respondents who are representative of the eventual test population. Prior to pretesting the open-ended questions, the test officer needs to be sensitive to the phrasing of the questions since inadvertent phrasing of the open-ended questions can sometimes modify responses in unrecognized and unintended ways. The use of open-ended response formats and interviews should enable the formulation of a questionnaire to obtain evaluative information. These interviews will provide guidance to the formulation of a sound survey instrument in the following respects:

(1) A better understanding of the factors or criteria which make up the mental set of individuals in evaluating systems and equipment.

(2) Some idea of the range of favorable and unfavorable opinions toward the system for each factor.

(3) Tentative knowledge of individual and group differential opinions toward the system tested.

Therefore, before drafting the pretest questionnaire, the researcher must have a feel for: question categories (e.g., problem areas, positive aspects); response categories (e.g., evaluative factors); and the type of system operations information which is needed (e.g., In evaluating a new helmet suspension system, does respondent wear eyeglasses?).

3. Using the Critical Incident Technique to Determine Questionnaire Content

The critical incident technique consists of a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness either in solving practical problems or in developing broad psychological principles. The technique calls for collecting observed incidents of behavior that have special significance and meet systematically defined criteria. It can be of assistance, therefore, in helping to determine the content of items to be included in a questionnaire.

Although there are a number of variations in the critical incident technique, the basic procedure consists of collecting records of specific behaviors related to the topic of concern. The behaviors might be noted by observers, or individuals can be asked to recall and record past specific behaviors judged to provide significant or critical evidence related to the topic of concern. As appropriate, behaviors related both positively and negatively to the area of concern should be noted. The records of behavior that are collected can then be analyzed and used as a basis for determining questionnaire content.

One of the examples of the use of the critical incident technique reported by Flanagan in the articles noted in Section III-B 3, had to do with a study of combat leadership in the United States Army Air Forces in 1944. It represented "the first large-scale, systematic effort to gather specific incidents of effective or ineffective behavior with respect to a designated activity. The instructions asked the combat veterans to report incidents observed by them that involved behavior which was especially helpful or especially inadequate in accomplishing the assigned mission. The statement finished with the request, 'Describe the officer's action. What did he do?' Several thousand incidents were collected in this way and analyzed to provide a relatively objective and factual definition of combat leadership. The resulting set of descriptive categories was called the 'critical requirements' of combat leadership" (p. 328).

For more information on the critical incident technique, see, for example, the following two sources:

a. Barnes, T. I. (1960). The critical incident technique. Sociology and Social Research, 44, 345-347.

b. Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-358.

4. Using Impressions of a Topic to Determine Attitude Scale Content

When the questionnaire is an attitude scale, a useful method for selecting items for it is to ask a group of individuals to write six statements giving their impressions of a topic, such as Army pay. From these, some smaller number of statements can be selected that are readable, intelligible, and capable of classification. These statements can then be sorted into several categories, such as the status of the topic and its good and bad features.

C. Other Considerations Related to Questionnaire Content

This section discusses a number of topics related to questionnaire content: questions that should be asked related to questionnaire content; sources of bias in questionnaire construction; and characteristics of good questions that affect questionnaire content.

1. Questions That Should Be Asked Related to Questionnaire Content

Asking yourself the following five questions may lay the foundation for a far more valuable questionnaire than would otherwise be produced. If you can't answer these questions, be sure to read or re-read the Outline Test Plan and the Independent Evaluation Plan.

a. Who needs the information? Knowledge of who needs the information will provide a source in the event answers are needed to the following four questions.

b. What decisions will be made based on your information? This will tell in part why the information is needed. Depending on what decision is going to be made, some kinds of information will make a difference and should be collected, and other kinds will not.

Suppose, for example, information is to be collected as a part of a test comparing a new item of equipment with an old standard item. The nature of the decision to be made is clear enough. It will be either selection of the new equipment, or retention of the old with which it is being compared. The basis for the decision will usually also be clear from the small development requirement (SDR) or qualitative materiel requirement (QMR) which led to the development of the item being tested. Analysis of the QMR will identify the qualitative requirements the new equipment must have, and will give the start needed to develop questions.

c. What facts will affect the decision? While this may be a difficult question to answer, trying to do so should identify items of information that should be sought with the questionnaire. It may also head off the collection of unnecessary information.

d. Whom are you asking? To get good information, not only must a good question be asked, but it must be asked of someone who has the answer. It would not, for example, be reasonable to ask support troops in a supply depot questions about combat operations.

e. What are the consequences of a wrong answer? While this basically is an administrative question, it has an important bearing on field questionnaire design. Clearly, if it makes little difference which of two alternatives is chosen, it makes little difference if the information is collected. On the other hand, if there is a chance that substantial dollar savings will result from the use of a more effective training technique, or that millions of dollars will be wasted by buying a new piece of equipment which is not better than the old, it is necessary to design tests very well, and ask the right questions with great care.

2. Refining Questions

Early versions of questions usually need to be refined. The following approaches will assist in developing better questions:

a. Try out questions on co-workers.

b. Identify problems in question wording prior to pretesting.

c. Pretest the questionnaire, and modify as needed. This should help in making the questionnaire easier for the respondents to use, and to assure meeting the objectives of the field test.

3. Sources of Bias in Questionnaire Construction

Two primary sources of bias in questionnaire construction that have been identified are investigator bias and question bias.

a. Investigator bias arises from: choice of subject matter; study design and procedure; unfair or loaded phrasing of questions; and interpretation and reporting of results. Sources of such biases include: the questionnaire developers' relationships with the clients; their personal involvement in a particular theoretical position or research technique; and those personal traits attributable to class, race, or political ideology. To reduce the impact of such bias, questionnaire developers need to: be aware of the problems; seek critiques from independent sources; carefully review previously published related reports; and continue pursuing technical improvement in their investigations.

b. Four ways that have been suggested of minimizing question bias when asking opinion questions are: ask many questions on the same topic; determine by scale analysis whether questions ask the respondents about the same dimensions of opinion (see Chapter V); ask "How strongly do you feel about this?" after each opinion question; and relate the content of opinion to the intensity of feeling.

Chapter IV: Types of Questionnaire Items

A. Overview

This chapter discusses various types of questionnaire items: open-ended items (Section IV-B), multiple choice items (Section IV-C), rating scale items (Section IV-D), behavioral scale items (Section IV-E), ranking items (Section IV-F), forced choice and paired-comparison items (Section IV-G), card sorting items/tasks (Section IV-H), and semantic differential items (Section IV-I). For each of these major item types, definitions and examples are presented, advantages and disadvantages are noted, and recommendations regarding their use in Army field test evaluations are given. Other types of items are noted in Section IV-J: checklists, matching items, arrangement items, and formats providing for supplementary responses.

It may be noted that a number of ways have been utilized in the professional literature for differentiating and classifying item types. Which types are special cases of other types could be debated at length. Unanimous agreement with the definitions given in this manual cannot, therefore, be anticipated.

B. Open-Ended Items

1. Definition and Examples

Open-ended items are those which permit respondents to express their answers to the questions in their own words, and to indicate any qualifications they wish. They are like general questions asked in an unstructured interview. By contrast, in a closed-end item, all the answers/choices/responses permitted are displayed, and respondents need only to check their preferred choices. Examples of open-ended items are shown in Figure IV-B-1.

Figure IV-B-1

Examples of Open-Ended Items

1. Describe any problems you experienced in moving through the test course while wearing the new PRC-99 radio harness.

   _____

   _____

2. The M16 rifle is: _____

   _____

3. What do you think of the AR-15 rifle sight? _____

   _____

2. Advantages of Open-Ended Items

a. Questions with open-ended response formats allow the respondents considerable latitude in their responses.

b. Open-ended items allow for the expression of middle opinions that closed-end items with two choices would not.

c. Open-ended items allow for the expression of issues of concern that may not have been identified by the question writer.

d. Open-ended items allow researchers to obtain answers that are unanticipated; unique information may be provided.

e. Open-ended items are very easy to ask. This is useful when the question writer either does not know, or is not certain about, the range of possible alternative answers.

f. With an open-ended question, it is possible to find out what is salient to the respondents, what their frame of reference is, and how strongly they feel.

g. Open-ended questions permit respondents to describe more closely and fully their real views.

h. There are times when more valid answers may be obtained from open-ended than closed-end items. For example, there may be a tendency for respondents to inflate yearly income figures. Providing response alternatives may result in an even greater inflation.

i. Answers to open-ended questions may be useful when treated as anecdotal material.

j. Respondents like the opportunity to answer some questions in their own words.

3. Disadvantages of Open-Ended Items

a. Open-ended items are time consuming for the respondent.

b. Open-ended questions which are self-administered and/or group-administered place a burden on the reading and writing skills of the respondent.

c. Asking people to answer questions in their own words increases the task difficulty, and can affect the rate of response. For example, respondents may say that they have no problems rather than taking the time to write out what the problems are. Item 1 in Figure IV-B-1 is poor in this respect, but item 2 is worse.

d. Only highly motivated respondents will take the time to write a complete answer to each question.

e. Open-ended items often leave the respondents on their own to determine what is relevant in the evaluation. For instance, item 2 in Figure IV-B-1 leaves the respondents to determine what is relevant in evaluating the M16 rifle. This is inappropriate. Open-ended questions should not be used to bypass the understanding of operations that the questionnaire writer should have or should acquire before preparing the final version of the questionnaire.

f. Questionnaires that use closed-end items are generally more reliable than those using open-ended items.

g. Open-ended questions, answered by motivated respondents, are capable of overloading data analysts. They usually cannot be handled by machine analysis methods without lengthy preliminary steps. Analysis of the responses to an open-ended question usually must be done by someone who has substantial knowledge about the question's content, rather than by a statistical clerk. They are often difficult to code for analyses. Thus, the data analysis task can grow into a major project and problem.

h. Open-ended questions may be easier to misinterpret since the respondent does not have a set of response alternatives available which might in themselves provide the proper frame of reference.

i. Much of the material obtained from an open-ended question may be repetitious or irrelevant.

j. Since open-ended questions are more time consuming, a constraint is placed on the number of questions that can be asked.

k. Open-ended questions are more subject to interviewer variations than are closed-end questions.

l. Open-ended items are often harder for the respondent to answer than closed-end questions. For example, respondents, when asked their annual income, may have to struggle to come up with relatively specific figures, whereas when response alternatives are presented, they need only indicate one of a number of ranges of income.

m. Inadvertent phrasing of open-ended questions can sometimes modify responses in unrecognized and unintended ways. It is difficult to predict in advance which words will bias an item. Subtle words appear to cause more distortion than blatantly biasing words.

4. Recommendations Regarding Use

a. Open-ended questions should be rarely used and, even then, such questions should sharply focus respondents' attention and thereby reduce their writing burden.

b. Closed questions are better for self-administered questionnaires than open questions.

c. In situations where time and money constraints are paramount, it would be more appropriate to use closed questions.

d. Closed questions are preferred for surveys where the responses would more likely be dichotomous.

e. For collecting nominal data, the researcher has a choice about whether to ask open-ended or closed-end questions.

f. When responses can be obtained by degree (for example, strongly agree to strongly disagree), a closed-end question would be superior to an open-ended question.

g. Sometimes a good procedure is to use an open-ended question with a small number of respondents as a pretest, in order to find out what the range of alternatives is. It may then be possible to construct good closed-end questions that will be faster to administer and easier to analyze.

h. Open-ended questions are most useful when there are too many possible responses to be listed or foreseen; when it is important to measure the saliency of an issue to the respondent; or when a rapport-building device is needed in an interview.

i. To obtain in-depth information on various content areas, a more focused and guided approach would be the use of an interview with open questions.

j. Use long open questions with familiar wording for questions with potentially threatening content.

k. It is sometimes useful to include one or more open-ended questions along with closed-end questions in order to obtain verbatim responses or comments that can be used to provide "flavor" of responses in a report.

C. Multiple Choice Items

1. Definition and Examples

In a multiple choice item, the respondent's task is to choose the appropriate or best answer from several given answers or options. As used here, multiple choice items include dichotomous or two-choice items as special cases. And, since only the permitted answers are available for selection, the multiple choice item may also be termed a closed-end item.

Examples of multiple choice items are shown in Figure IV-C-1. Items 3, 4, and 5 are dichotomous, i.e., provide two response alternatives.

A comparison of true-false items with nondichotomous multiple choice items is made in Section VI-G, since they are issues related to the number of response alternatives.

2. Advantages of Multiple Choice Items

a. As seen in item 2 of Figure IV-C-1, the questionnaire writer may select different numbers of response alternatives depending upon knowledge of the respondent's experience or depending upon the decision to allow or disallow respondents to "sit on the fence" by including a "no preference" alternative. (See Section VI-C for wording of items, and Section VI-G regarding the number of response alternatives to employ.)

b. Responses are more reliable when response alternatives are provided for respondents.

c. Interpretation of responses is more reliable when response alternatives are provided to respondents.

d. Dichotomous items are relatively easy to develop, and permit rapid analyses.

e. Complex questions can often be broken down into two or more simpler questions.

f. Multiple choice items are easily scored, which means that data analysis is a relatively inexpensive process requiring no special content expertise.

g. Multiple choice items require considerably less time per respondent answer than open-ended items.

h. Multiple choice items put all persons on the same footing when answering. That is, each person will be able to consider the same range of alternatives when choosing an answer.

i. Multiple choice items are easy to administer.

Figure IV-C-1

Examples of Multiple Choice Items

1. What do you consider the most important characteristic of a good helmet? (Check one)

   _____ Comfort

   _____ Stability

   _____ Utility for wash basin

   _____ Protection

   _____ Weight

2. Which dc you prefer, the M16 or the M14 rifle? (Check one)

   _____ M14

   _____ M16

   _____ No preference

3. Were you able to fire effectively from the frontal parapet emplacement?

   _____ Yes _____ No

4. Which do you prefer, the ABC helmet or the XYZ helmet?

   _____ ABC helmet _____ XYZ helmet

5. The M16 is a better rifle than the M14.

   _____ True _____ False

6. What is your marital status?

   _____ Single

   _____ Married

   _____ Divorced

   _____ Other (e.g., separated, widowed, etc.)

3. Disadvantages of Multiple Choice Items

a. Dichotomous items force the respondents to make a choice even though they may feel there are no differences between the alternatives, or they do not know enough about either to validly choose one. Furthermore, respondents are not permitted to say how much better one alternative is than the other.

b. Two alternatives might not be enough for some types of questions. The question designer may oversimplify an issue by forcing it into two categories.

c. There may be a tendency for respondents to choose an answer on the basis of a response set. (See Chapter XII.)

d. Unless care is taken in the construction of multiple choice items, the response alternatives may overlap.

e. The question maker has to know the full range of significant possible alternatives at the time the multiple choice question is formulated.

f. Multiple choice items must be worded with very great care. Otherwise, the information obtained may not be valid.

g. With dichotomous items, any slight language difficulty or misunderstanding of even one word could change the answer from one extreme to another.

4. Recommendations Regarding Use

a. For some purposes, the dichotomous question (two response alternatives) may be an improvement over the open-ended question in that it provides for faster and more economical analysis of data. However, it requires more care in its development.

b. Generally speaking, dichotomous multiple choice questions should be avoided. If used, they should probably be followed-up to determine the reason for a given response.

c. Nondichotomous multiple choice items are popular and have wide utility. They are recommended for general use as appropriate.

d. Forced response and multiple choice items are desired when measuring soft data such as opinions. Checklists are recommended for hard data such as physical aspects of a job analysis or a broad generalization for measuring opinions prior to a later survey.

e. The development of questionnaire items should include pilot testing using open-ended items which are later converted to multiple choice items.

f. No one scaling format has consistently been superior to another. Rating scales need to be evaluated on other criteria than number of scale points, vertical and horizontal formats, and unipolar or bipolar scales.

g. Prior to multiple choice format selection, the type of measurement scale and data analysis should be identified.

h. Multiple choice items represent measurement scales which are nominal, ordinal, or interval. These measurement categories indicate the rules for assigning numbers to the data so that the appropriate statistical analyses can be performed.

i. Ordinal measurement scales are common in surveys where respondents are required to rank items or to use a paired-comparison method.

j. One item cannot adequately cover a topic area. It is necessary to develop many items to avoid obtaining only surface facts, and to provide the researcher with a deeper understanding of the relevant experience of the respondents.

k. Multiple choice items can be developed which measure higher order objectives.

l. If multiple questions are asked about different possible responses to a problem, separate specific questions that can be understood by all respondents and easily interpreted are required.

m. The length of an item may possibly modify the response style. Researchers may wish to develop alternate versions of questionnaire items where the different versions are of different lengths. This would allow comparison of the effect of item length on responses.

D. Rating Scale Items

1. Definitions and Examples

Rating scale items are a variation of multiple choice items. They are a means of assigning a numerical value to a person's judgment about some object. They call for the assignment of responses either along an unbroken continuum or in ordered categories along the continuum. The end result is the attachment of numbers to those assignments. Ratings may be made concerning almost anything, including people, groups, ourselves, objects, and systems.

There are a number of different forms of rating scale items, only two of which are shown here. Figure IV-D-1 shows examples of "numerical" scales. In item 1, a sequence of defined numbers is provided for the respondent.

Figure IV-D-1

Examples of Numerical Rating Scale Items

1. The cleaning kit for the M16 rifle is

_____ 7 very easy to use.
_____ 6 quite easy to use.
_____ 5 fairly easy to use.
_____ 4 borderline.
_____ 3 fairly difficult to use.
_____ 2 quite difficult to use.
_____ 1 very difficult to use.

2. How satisfied or dissatisfied are you with the type of furniture in the barracks?

_____ Very satisfied
_____ Satisfied
_____ Borderline
_____ Dissatisfied
_____ Very dissatisfied

3. The training that I have received at Fort Hood has been

_____ very challenging.
_____ challenging.
_____ borderline.
_____ unchallenging.
_____ very unchallenging.

The respondents are to indicate which defined number best fits their judgment about the object to be rated. Sometimes, the numbers are not shown on the form used by the respondent (e.g., items 2 and 3). Instead, the respondent reports in terms of descriptive cues and the numbers are attached later during analysis. The numbers assigned are in an arithmetic sequence, such as 5, 4, 3, 2, 1, depending upon the number of response alternatives used. They are usually assigned arbitrarily unless the response alternatives have been scaled using one of the procedures described in Section V-B. The order of perceived favorableness of commonly used words and phrases is discussed in Chapter VIII.

Figure IV-D-2 shows an example of a graphic rating scale. In the graphic scale, the descriptors are associated with points on a line or graph, and the respondent indicates a judgment by marking the point on the line which best fits the rating of the object. The line can be either horizontal or vertical. The graphic scale allows the respondent to place a judgment any place on the line. Thus, the respondents are not confined to discrete categories as they are with the numerical scale. It is, however, more difficult to score, but this can be facilitated with a stencil which divides the line into segments to which numbers are assigned.

The number of response alternatives to use is discussed in Section VI-G, the order of response alternatives in Section VI-H, and response anchoring in Chapter VII.

Figure IV-D-2

Example of Graphic Rating Scale Item

1. Place an X at the point on the scale that most clearly represents your opinion about the cleaning kit for the M16 rifle.

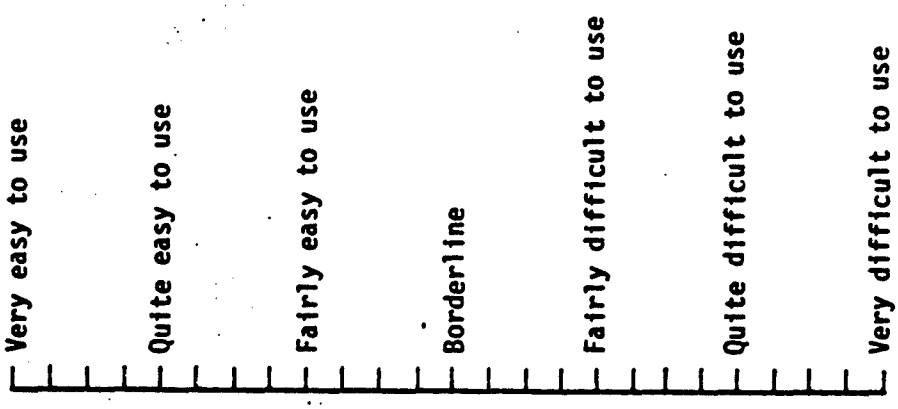
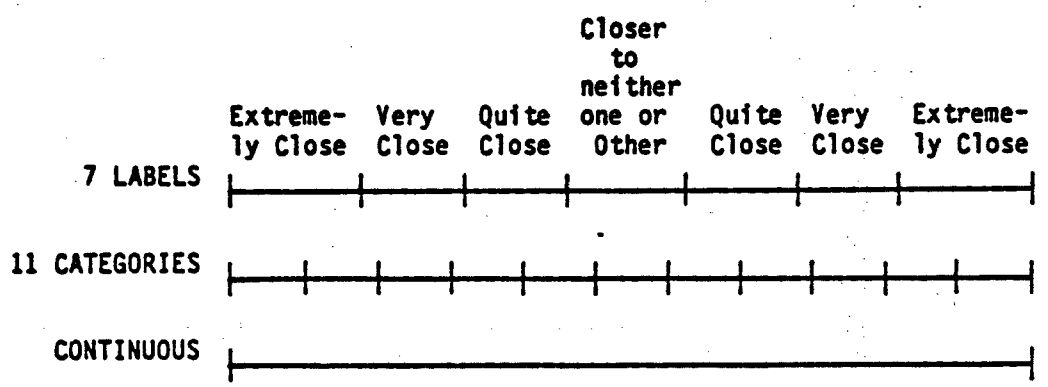


Figure IV-D-3 shows examples of continuous scales.

Continuous scales are usually thought of as straight lines with no indications of any differentiation along the scale lines. A continuous scale can provide the respondent with guidance as to the directionality of the rating, and offer the respondent greater discrimination as to ratings along the scale line. Continuous scales have been used in ergonomics to rate perception of a thermal stimulus as well as to rate perception of tones.

Figure IV-D-3

Examples of Discrete and Continuous Scales
Used to Rate Perception of Tones



2. Advantages of Rating Scale Items

a. When properly constructed, the rating scale reflects both the direction and degree of attitude or opinion, and the results are amenable to analysis using conventional statistical procedures.

b. Graphic rating scales allow for as fine a discrimination as the respondent is capable of giving, and the fineness of scoring can be as great as desired.

c. Rating scale items usually take less time to answer than do other types of items.

d. Rating scale items can be applied to almost anything.

e. Continuous scales may at times yield greater discrimination by raters.

f. Rating scale items are generally more reliable than dichotomous multiple choice items. They may be more reliable than paired-comparison items.

g. Manipulation of the anchors does not appear to greatly affect the results. The inadvertent use of mismatching antonyms with partial antonyms to anchor a rating scale may not jeopardize the reliability of the scale.

3. Disadvantages of Rating Scale Items

a. Rating scale items are more vulnerable to biases and errors than other types of items such as forced choice items.

b. Graphic rating scales are harder to score than other types of items. With a graphic scale item format, the verbal anchors are associated with points on a line, and the respondents indicate their judgment by marking the point on the line which best represents their judgment. Considerable effort and time are required to measure the pencil mark's exact location to the nearest portion of the line.

c. The results obtained from the use of graphic rating scale items may imply a degree of precision/accuracy which is unwarranted.

4. Recommendations Regarding Use

a. The use of rating scale items is highly recommended for most questionnaires.

b. Rating scales present the sentence (stem) first, and require the respondent to select a response alternative to complete the sentence. The stem is supposed to be neutral so that the response alternatives contain different combinations of directionality (positive or negative) and intensity.

c. Scales having apparently equal intervals should be employed. The respondent will assume or perceive that the distances between adjacent scale points are equal.

d. Numbers can be presented along with verbal anchors.

e. Applications which require greater discrimination could use scales with more than five or six categories, or with continuous lines.

f. It is possible to develop and apply a continuous scale without affecting the psychometric properties of the scale. Continuous scales appear to be equivalent to traditional scales with discrete categories.

g. Minor violations in the technique of scale development for bipolar anchors, such as quasi-polar anchors and phrases for anchors, do not appear to threaten the reliability of the instrument. Therefore, it is possible to establish new versions for bipolar anchors.

E. Behavioral Scale Items

1. Definition and Examples

Behavioral scale items are derived from the compilation of critical incidents (whether really critical or not). They were developed to encourage raters to observe behavior more accurately. Behavioral scales have evolved using different developmental procedures with divergent scaling foundations associated with Likert, Thurstone, and Guttman scales. There are a variety of behavioral scales such as Behaviorally Anchored Rating Scales (BARS), Behavioral Expectation Scales (BES), Behavioral Observation Scales (BOS), and Mixed Standard Scales (MSS).

Behavioral scales have customarily been used to evaluate individual performance on the job. There have been other applications that include assessing morale, and a tool to make decisions about the effectiveness of maintenance trainer equipment and actual equipment training.

Even though developmental procedures vary according to the type of behavioral scale, there are some commonalities. Behavioral scales are built on large numbers (in the hundreds) of critical incidents which are reduced in number by being fitted into performance dimensions and/or categories. There must be a specified level of agreement (usually somewhere between 60% and 80%) to retain a critical incident for inclusion in the scale. The critical incidents are anchored to the scale. Critical incidents describe a continuum of effective and ineffective behavior.

Procedures for constructing behavioral scale items, and evaluative comments about them, can be found in a number of sources including the following:

a. Bernardin, H. J., & Smith, P. C. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales. Journal of Applied Psychology, 66(4), 458-463.

b. Borman, W. C. (1979). Format and training effects on rater accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

c. Katcher, B. L., & Bartlett, C. J. (1979, April). Rating errors of inconsistency as a function of dimensionality of behavioral anchors (Research Report No. 84). College Park, MD: University of Maryland, Department of Psychology. (DTIC No. AD A068922)

d. Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 34, 263-289.

e. Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3(2), 193-200.

f. Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. Personnel Psychology, 32, 299-311.

g. Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62(2), 177-183.

h. Murphy, J. W. (1980). Use of behaviorally anchored rating scales (BARS) to complement the management by objectives (MBO) and fitness report components of the Marine Corps performance evaluation system. Master of Military Arts and Sciences (MMAS) thesis prepared at U.S. Army Command and General Staff College, Fort Leavenworth, KS. (DTIC No. AD A097694)

Examples of behavioral scale items and dimensions are shown for BARS, BES, BOS, and MSS in Figures IV-E-1 through IV-E-4.

Figure IV-E-1

Examples of BARS's Seven Dimensions
Describing Technician Behavior

1. Safety: Behaviors which show that the technician understands and follows safety practices as specified in the technical data;

2. Thoroughness and Attention to Details: Behaviors which show that the technicians are well prepared when they arrive on the job, carry out maintenance procedures completely and thoroughly, and recognize and attend to symptoms of equipment damage or stress;

3. Use of Technical Data: Behaviors which show that the technician properly uses technical data in performance of maintenance functions;

4. System Understanding: Behaviors which show that the technicians thoroughly understand system operation allowing them to recognize, diagnose, and corrrect problems not specifically covered in the Technical Orders and publications;

5. Understanding of Other Systems: Behaviors which show that the technicians understand the systems that are interconnected with their specific system and can operate them in accordance with technical orders;

6. Mechanical Skills: Behaviors which show that the technician possesses specific mechanical skills acquired for even the most difficult maintenance problems; and

7. Attitude: Behaviors which show that the technician is concerned about properly completing each task efficiently and on time.

From Wienclaw, R. A., & Hines, F. E. (1982, November). A model for determining cost and training effectiveness trade-offs. Training Equipment Interservice/Industry Training Equipment Conference, 405-416.

Figure IV-E-2

Example of BARS Items Representing
Performance and Effort on the Job

Scale Point	Behavioral Anchor
9	When maintenance mechanics found an error in their assembly procedures on an aircraft, they told their platoon leaders of their mistake and requested that the hangar be open Saturday and Sunday if necessary to meet their previously promised Monday delivery.
8	While clearing the brush from an approach to an airport, these dozer operators never shut the dozer off, running in shifts right through lunch.
7	This section was asked to prepare a set of firing charts by a specific time. The charts were finished ahead of time.
6	Although this section was constantly called upon for typing tasks, the work was done with few mistakes and on a timely basis.
5	The men in this unit did not push for top performance, although they did their jobs and kept busy.
4	Many troops in this unit would leave the post as quickly as possible after duty hours to avoid doing any extra work.
3	The service section of a support unit had a large backlog of equipment needing repair. All enlisted personnel assigned to this section appeared to be busy, but their output was very low compared to the other service sections.
2	The men in this section signed out weapons to be cleaned but sat around and "shot the bull" until it was time to turn the weapons back in.
1	During one period, these enlisted personnel slowed their work down and made mistakes that cost time and new parts. They were working 7-day weeks, but at the end of the period, they were accomplishing only the same amount of work in 7 days that they had been accomplishing before in 5 days.

From Motowidlo, S. J., & Borman, W. C. (1977). Behaviorally anchored scales for measuring morale in military units. Journal of Applied Psychology, 62(2), 177-183.

Figure IV-E-3

Example of BOS Item Representing
Description of Foreman's Job

Tells crew to inform him immediately of any unsafe condition.

Almost Never 1 2 3 4 5 Almost Always

From Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foremen. Personnel Psychology, 32, 299-311.

Figure IV-E-4

Example of MSS Items Representing
Highway Patrol Stopping Vehicles for Violations

- Stops vehicles for a variety of traffic and other violations.
- Concentrates on speed violations, but stops vehicles for other violations also.
- Concentrates on one or two kinds of violations and spends too little time or others.

From Rosinger, G., Myers, L. B., Levy, G., Loar, M., Mohrman, S. A., & Stock, R. (1982). Development of behaviorally based performance appraisal system. Personnel Psychology, 35, 75-88.

2. Advantages of Behavioral Scale Items

   a. Raters may not be cognitively prepared to summarize and abstract accurately. More reliable ratings may be obtained on behavioral scales by using the jargon of raters, and by having raters maintain observational diaries.

   b. It has been found that it is possible to generalize a Behaviorally Anchored Rating Scale (BARS) instrument for use with similar populations in other organizations where the same types of tasks are being performed.

   c. Behavioral Expectation Scales (BES) can be used to clarify organizational policy, provide feedback, assess and improve individual performance, and identify divergent perceptions.

   d. Training programs of three hours and longer have the potential to increase rater accuracy.

   e. In situations where there is concern about halo and leniency errors, Mixed Standard Scales (MSS) would be appropriate to use if the developmental procedures are thorough.

3. Disadvantages of Behavioral Scale Items

   a. The time and effort involved in developing behavioral scale items may not be worth the investment unless there are other spin-offs for the use of this type of scale.

   b. Behavioral scales require quantification of items using a sample size of several hundred people; they should not be based on small samples.

   c. More items are generated for behavioral scales when the number of dimensions is increased. For example, there is the potential for nine dimensions to have up to 90 items or more.

   d. Raters appear to prefer a BARS format over a MSS format. It would probably not be useful to construct a MSS unless halo and leniency errors were anticipated.

4. Recommendations Regarding Use

a. Scale development procedures will be strengthened if rater participation is included for BARS as well as other behavioral scale formats.

b. BARS development procedures have resulted in a disproportionate rejection of mid-range items. Simple item intercorrelation procedures for the U_c (universe score procedure) would increase the number of mid-range items. (DeCotiis, T. A. (1978). A critique and suggested revision of behaviorally anchored rating scales developmental procedures. Educational and Psychological Measurement, 38, 681-690.)

c. Rigor in the developmental procedures for constructing various types of behavioral scales will influence and increase the reliability and validity of the scales more than the format.

d. There appears to be a tendency to confound Thurstone scaling procedures with Likert scaling procedures which diminishes levels of reliability and validity for Thurstone scales. Researchers need to be aware of the differences between Thurstone and Likert scale development procedures when they are constructing BARS, BES, and BOS behavioral scales.

e. To increase the MSS format acceptance by raters for the scoring system and item dimensionality, a coding system with face validity may be useful as well as training for the raters to explain the MSS rationale, and the procedures for carrying out the appraisal.

f. MSS requires statistical analysis to ensure unidimensionality of the scales.

F. Ranking Items

1. Definition and Examples

Ranking items call for the respondent to indicate the relative ordering of the members of a presented group of objects on some presumably discriminable dimension, such as effectiveness, saltiness, overall merit, etc. By definition, one does not have a scale by which the amount of difference between successive members is measured, nor is it implied in rank ordering that successive differences are even approximately equal. If respondents were being asked to give judgments on the size of intervals, the item would be something more than a ranking item.

Multiple choice items are so frequently used that one may inadvertently use this format when the ranking item format would provide more complete and reliable information. Item 1 in Figure IV-C-1 illustrates this point. Since a preponderance of respondents would check "protection" as a helmet's most important characteristic, only a small remainder of responses would be available as a basis for ordering the other characteristics. Some of the other characteristics might be achievable without sacrificing protection, so it would be desirable to have a reliable ordering of their importance.

As the number of objects to be ranked increases, the difficulty of assigning a different rank to each object increases even faster. This means that reliability (repeatability) is reduced. To counter this, one may explicitly permit respondents to assign tied rankings to objects when the number of objects exceeds, say, 10 or more.

Examples of ranking items are shown in Figure IV-F-1.

There have been instances when rank order scaling procedures have been integrated with other complex systems. An illustration of this is the delta scalar method used by the U.S. Navy and the Air Force Aerospace Medical Research Laboratory. The delta scalar method is a complex system of rank ordering found in the Mission Operability Assessment Technique and Systems Operability Measurement Algorithm (U.S. Navy), and the Subjective Workload Assessment Technique (U.S. Air Force). These systems involve establishing a rank order scale that is converted to an interval scale. Procedures and recommendations for constructing rank ordering embedded in subjective workload assessment methods can be found in a number of sources including:

a. Eggemeier, F. T., Crabtree, M. S., & La Point, P. A. (1983, October). The effect of delayed report on subjective ratings of mental workload. Proceedings of the Human Factors Society 27th Annual Meeting, 139-143.

b. Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., & Shingledecker, C. A. (1982). Subjective workload assessment in a memory update task. Proceedings of the Human Factors Society 26th Annual Meeting, 643-647.

c. Eggemeier, F. T., McGhee, J. Z., & Reid, G. B. (1983, May). The effects of variations in task loading on subjective workload rating scales. Proceedings of the IEEE 1983 National Aerospace and Electronics Conference, Dayton, OH, 1099-1105.

Figure IV-F-1

Examples of Ranking Items

1. Rank the following three methods of issuing starlight scopes to an infantry squad. Assign a "1" to the most effective, a "2" to the second most effective, etc. Do not assign tied rankings.

Ranking	Basis of Issue
_____	Scopes issued to AMG and SL
_____	Scopes issued to AMG, SL, and one rifleman
_____	Scopes issued to all squad members

2. How important are each of the following factors to you? Assign a "1" to the most important, "2" to the second most important, etc. Assign a different number to each of the four factors.

_____ Type of furniture in the barracks

_____ Army pay

_____ Medical service to soldiers

_____ Choice of duty station

2. Advantages of Ranking Items

a. The idea of ranking is familiar to respondents.

b. Ranking takes less time to administer, score, and code than paired-comparison items do, and there is some evidence that the results of the two are highly similar.

c. Ranking and rating techniques are generally comparable in terms of reliability.

3. Disadvantages of Ranking Items

a. Ranking items such as item 1 in Figure IV-F-1 do not reveal the respondent's judgment as to whether any of the objects are effective or ineffective in an absolute rather than just a relative sense. To learn this, another question must be asked.

b. Rank order scales originate from ordinal scale measurement. The categories in a rank order scale do not indicate how much distance there is between each category. Unequal distances are assumed. Rank order items do not permit respondents to state the relative amounts of differences between alternatives.

c. The results from ranking items are open to question if the basis for ranking was not clear to the respondents.

d. Ranking is generally less precise than rating.

4. Recommendations Regarding Use

a. Rank order scales are appropriate for analyzing data that meets the requirements of ordinal measurement scales.

b. There are some situations where the intent of the questionnaire developer is best served with the use of one or more ranking items. Generally, however, rating scale items are probably preferable.

c. Rank order scales and rating scales are more cost effective and time effective to use than paired-comparisons.

d. Individuals tend to more frequently use one end of a list than the other end while ranking. To counteract this bias, it is possible to develop two or more versions of the list by randomly ordering the lists.

e. It is possible to combine rank ordering with other methods, such as task analysis, to isolate critical components of a job. This information can be transformed into a performance measurement system, or can be used to modify military training.

f. Analysis of the data for test-retest reliability performed on rank order, paired-comparison, and Likert scales varied depending on whether a Spearman rho or Kendall's tau was used. Kendall's tau may be a more appropriate measure of reliability for rank order measures.

G. Forced Choice Items

1. Definition and Examples

It would appear that any multiple choice item could also be called a "forced choice" item because, after all, the respondent is expected to choose one of the response alternatives. The instructions and/or the presence of an administrator put some degree of social pressure - social force - on the respondent. However, if a multiple choice item includes an "I don't know" response alternative, the pressure/force is almost totally removed. Likewise, on a rating scale item, the inclusion of a "neutral" or "borderline" response category allows the respondents to answer without committing themselves.

So, for some questionnaire developers - in particular those who produce "forced choice self inventories" (see references) - a "forced choice" item strictly refers to one where the respondents must commit themselves. They may have to select one of a pair of choices, or two of three, or two of four. These three cases are illustrated in Figure IV-G-1.

2. Advantages of Forced Choice Items

a. Studies have indicated that reliabilities and validities obtained from the use of forced choice items compare favorably with other methods.

b. The forced choice method has been used by a number of investigators in an attempt to control the tendency of individuals to answer self-report inventories in terms of response sets rather than giving "true" responses. (Response sets are discussed in Chapter XII.)

3. Disadvantages of Forced Choice Items

a. Respondents sometimes balk at picking unfavorable statements, or at being forced to make a choice.

b. Forced choice items take more time to develop than some other types of items.

c. Paired-comparison items, where all phrases are paired, take more time to administer, score, and code than do ranking items. Results from the two, however, may have a linear relationship.

Figure IV-G-1

Examples of Forced Choice Items

1. Check one of the following two statements that is more characteristic of what you like.

   _____ I like to travel.

   _____ I like to meet new people.

2. Check one of the two following statements that is more characteristic of yourself.

   _____ I am honest.

   _____ I am intelligent.

3. Look at the following three activities. Mark an "M" by the one you like the most, and an "L" by the one you like the least.

   _____ Play baseball

   _____ Go to the craft shops

   _____ Attend boxing or wrestling matches

4. From the following four statements, check the two that are most descriptive of your unit commander.

   _____ Serious-minded

   _____ Energetic

   _____ Very helpful

   _____ Gets along well with others

d. There is some question as to whether forced choice items overcome the biases or errors they are supposed to correct.

e. Some investigators have concluded that the generalization that self-report forced choice inventories are more valid than single stimulus forms of the same tests is not supported by a critical consideration of the relevant evidence.

Procedures for constructing forced choice items, and evaluative comments about them, can be found in a number of sources including the following:

a. Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.

b. Nunally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill, pp 484-485.

c. Sisson, E. D. (1948). Forced choice--the new Army rating. Personnel Psychology, 1, 365-381.

4. Recommendations Regarding Use

When test participants are deliberately given relevant experience with the operation of a weapons system, vehicle, or other system, the "I don't know" response alternative should normally be deleted from items that seek the participants' evaluations of the system.

H. Card Sorting Items/Tasks

1. Definition

With card sorting items/tasks, the respondents are given a large number of statements (e.g., 75), each on a slip of paper or card. They are asked to sort them into, say, nine or eleven piles. The piles are in rank order from "most favorable" to "least favorable" or "most descriptive" to "least descriptive," etc., depending upon the dimension to be used. Each pile usually is to have a specified number of statements placed into it as required to form a rough normal distribution. However, some investigators have argued that forcing a given distribution is not necessary. Ordinarily each pile is given a score value which is then assigned to the statements placed into it.

An extensive discussion of the use of card sorts (or, more generally, Q-technique and its methodology) appears in: Stephenson, W. The study of behavior. Chicago: University of Chicago Press, 1953.

2. Advantages of Card Sorting Items/Tasks

a. Card sorts appear to be capable of counteracting at least some of the biasing effects of response sets. (Response sets are discussed in Chapter XII.)

b. Some investigators believe that card sorting is a fast and interesting method of obtaining valid and reliable interview data.

c. With card sorts, the respondents can shift items back and forth if they wish to do so.

d. The card sort has greatest value when a comprehensive description by a single individual is desired.

e. Card sorts also have value for obtaining complex descriptions which can be compared systematically.

f. They can be used to obtain rating information on any issue.

3. Disadvantages of Card Sorting Items/Tasks

a. Card sorting items/tasks may take more time to construct than other types of items, and they generally take more time to administer and score.

b. Card sorts are more involved to administer than other types of questionnaire items.

4. Recommendations Regarding Use

Some authors think that card sorting is the method of choice if testing time is available. Its greatest value seems to be its ability to provide a comprehensive description by a single individual, or to obtain complex descriptions which can be systematically compared. Since it is more awkward to administer and score than other types of items, its use in Army field test evaluations is limited.

I. Semantic Differential Items

1. Definition and Examples

The semantic differential technique was initially developed as a general method of measuring meaning, and with it the meaning of a particular concept to a particular individual can be specified quantitatively. The technique has also been used to measure attitudes and values, particularly in the marketing area. In using the technique, the respondent is presented with a number of bipolar rating scales, usually but not always having seven points. The two ends of each scale are defined by adjectives. The respondent is given a set of such scales, and is asked to rate each of a number of objects or concepts on every scale. To aid in interpretation, some scale coding can be used, usually numbers in a direct numerical sequence such as 1 through 7. Other more extensive scoring can be used, and results can be factor analyzed to search for the basic dimensions of meaning. However, the usefulness of the semantic differential as a research tool stems from the ability of the procedure to probe into both the content and the relative intensity of respondents' attitudes.

Examples of semantic differential items are given in Figure IV-I-1. A recommended text on the semantic differential is Osgood C. E., Suci, G. J., & Tannenbaum, P. H. (1957). The measurement of meaning. Urbana, Ill., University of Illinois Press. Norms have been collected on 20 scales for 360 words. They are reported in Jenkins, J. J., Russell, W. A., & Suci, J. (1958). An atlas of semantic profiles for 360 words. American Journal of Psychology, 71, 688-699.

2. Advantages of Semantic Differential Items

a. Evidence on the validity, reliability, and sensitivity of the scales has been offered.

b. Using some adjectives that do not seem appropriate to the concept under investigation may uncover aspects that reflect an attitude or feeling tone even though the respondent cannot put it into words.

c. Semantic differential items can be used to study the relative similarity of different concepts to the respondent, and to study changes over time.

d. Semantic differential items are relatively easy to construct, administer, and score.

Figure IV-I-1

Examples of Semantic Differential Items

1. Place an X in each of the following rows to describe your assessment of the M16 rifle.

Reliable	___:	___:	___:	___:	___:	___:	___	Unreliable
Heavy	___:	___:	___:	___:	___:	___:	___	Light
Good	___:	___:	___:	___:	___:	___:	___	Bad
Slow	___:	___:	___:	___:	___:	___:	___	Fast
Adequate	___:	___:	___:	___:	___:	___:	___	Inadequate

2. Place an X in each of the following rows to describe your assessment of the ABC helmet.

Reliable	___:	___:	___:	___:	___:	___:	___	Unreliable
Heavy	___:	___:	___:	___:	___:	___:	___	Light
Good	___:	___:	___:	___:	___:	___:	___	Bad
Slow	___:	___:	___:	___:	___:	___:	___	Fast
Adequate	___:	___:	___:	___:	___:	___:	___	Inadequate

3. Disadvantages of Semantic Differential Items

   a. If care is not taken, the two adjectives chosen for the extremes will not define some kind of scale or dimension between them.

   b. The value of semantic differential items depends on the suitable choice of the bipolar adjectives and concepts.

   c. There is a potential response error present in the respondents' interpretations of the meaning of the end-point descriptions. However, there appears to be a balancing out over a number of administrations.

   d. There is the possibility of a socially desirable response set when personality traits are measured with the semantic differential.

4. Recommendations Regarding Use

a. There are a number of investigators that advocate the use of the semantic differential. Others, however, have questioned whether it may be a rather complicated way of developing a measure that is more readily and reliably secured by other means. It is reasonable to assume that the technique could easily be expanded to identify attitudes and the intensity of the attitudes toward the attractiveness of a particular military specialty, the capacities of a specific piece of equipment to perform, or any other characteristic set which can be described by bipolar adjectives. However, since the analysis of sets of semantic differential items is somewhat involved, the technique has not been widely used for routine Army field test evaluations.

b. Semantic space for the concepts of evaluation, potency, and activity are fairly stable across studies, and have maintained reliability over time. Because of the stability of the scale, it is possible to vary instrument format as well as rating instructions and maintain the viability of the scale. To ensure the soundness of the scale, developmental procedures need to include testing the instrument in the context area for which it was designed.

c. In the early stages of development for the semantic differential, it is possible to identify potential bipolar anchors using Roget's Thesaurus as a source in addition to the subjects' concepts of terms that have semantic stability. Initial pools of items can be reduced through judgment agreement, factor analysis, and cluster analysis.

d. Semantic differential scales can be anchored with phrases, adjectives, or adverbs.

e. The number of scale points used with the semantic differential can vary, and still retain the integrity of the instrument. An acceptable range in the scale would be between five and twelve points. Each completed survey would have all items with the same number of scale points. For example, two questionnaires could be designed, one with seven scale categories and the other with nine scale categories.

f. Social desirability response sets can be controlled by careful construction of the bipolar scales. Adjectives can be selected that reflect a common trait to control the influence of social desirability.

J. Other Types of Items

1. Checklists

Checklists are instruments in which responses are made by checking the appropriate statement or statements in a list of statements. Examples are shown in Figure IV-J-1.

Figure IV-J-1

Examples of Checklists

1. Which of the following are important to consider when deciding whether or not to make a career of the Army? Check all that apply.

   _____ Leadership of NCOs

   _____ Opportunity for promotion

   _____ Playboy magazines in the Post Exchange

   _____ Latrine in crafts shops

   _____ Army pay

   _____ Choice of duty stations

   _____ Civilian opinion of Army

   _____ Reenlistment bonuses

   _____ Hours of work in a work week

2. Please check all the characteristics which Backpack A possesses.

   _____ Durability

   _____ Lightness

   _____ Wearing comfort

   _____ Accessibility of items

   _____ Ease of putting on and taking off

   _____ Other (specify): _____

Checklists can be used in conjunction with interviews to serve as a cue to the interviewer. Administration of a checklist combined with an interview of critical areas identified on the checklist could reduce interviewing time. Examples are shown in Figure IV-J-2.

Figure IV-J-2

Example of Checklist Pertaining to Equipment Problems

I will name equipment from the LAVM/RV that you may have used to extract, replace and transport equipment. Please answer Yes or No to indicate whether or not you experienced any difficulties using the equipment. I would also appreciate your comments concerning the difficulties. If you have no experience using the equipment, then check Not Applicable (NA).

	Equipment	Yes	No	NA	Comment
1.	Crane	___	___	___	_____
2.	Crane remote controls	___	___	___	_____
3.	Crane onboard controls	___	___	___	_____
4.	Winch	___	___	___	_____
5.	Winch controls	___	___	___	_____

This checklist/interview could serve as the foundation for generating other, more refined instruments. The checklist/interview is another way of eliciting information from a subject matter expert group.

Compared to rating scales, which give a numerical value to some sort of judgment, checklists are relatively crude. They are, however, quite useful when scaled information is not needed. Checklists also are useful when information is needed to determine which of several issues are significant to a respondent. Other issues regarding the use of checklists are as follows:

a. Checklists should use terms like the respondent uses.

b. Response set can be somewhat controlled if the respondent is asked to check a stated number of items, or if upper or lower limits are set.

c. There is some evidence that a higher rate of claim or assertion is obtained from checklists than from open-ended items.

d. It is usually not known if checklists cover the appropriate attributes.

e. Adjective checklists are sometimes used, especially to elicit stereotypes about people or nations. They are similar to rating scales.

2. Matching Items

With matching items, the respondent is given two columns of items, and is asked to pair each item in the first column with an associated item in the second. In general, it is not desirable to have the same number of items in each column. Both sets of items should constitute a homogeneous set, and any item in the second column should look like it could go with any item in the first column.

Matching items are best used in achievement testing. Since they have little utility in Army field test evaluations, they are not discussed in greater detail in this manual.

3. Arrangement Items

With an arrangement item, a number of statements are presented in random order, and the respondent arranges them in a new order according to his/her judgment and the guidance received. For example, steps in a sequence of events or procedures may be rearranged in order of occurrence or performance. Or, causes may be rearranged in order of importance in bringing about a certain effect.

There may be some situations where arrangement items may be useful in Army field test evaluations; however, the scoring of the items is difficult. The use of such items is, therefore, extremely limited.

4. Formats Providing for Supplementary Responses

The questionnaire writer is not limited to the major item formats described in this chapter. Formats providing for supplementary responses can also be used. Examples are shown in Figure IV-J-3.

Figure IV-J-3

Examples of Formats Providing for Supplementary Responses

1. The starlight scope is able to detect aggressor movements:

   _____ very effectively.

   _____ effectively.

   _____ borderline.

   _____ ineffectively.

   _____ very ineffectively.

   Explain: _____

   _____

2. What style of leadership was used by the most effective squad leader you served under? (Check one)

   _____ democratic and friendly

   _____ friendly with most; authoritarian with the others

   _____ sometimes authoritarian; sometimes acts like one of the men

   _____ usually authoritarian; avoided making close friends

   _____ other (please describe) _____

   _____

Notice that the "other" response alternative in Example 2 allows the respondent in effect to make an open-ended item out of a multiple choice item. Few test respondents, however, elect to do this. Inclusion of the supplementary or write-in option commits you to extra data reduction and analysis effort that would have been unnecessary had you anticipated and included all reasonable response alternatives.

Chapter V: Attitude Scales and Scaling Techniques

A. Overview

At times, the questionnaire developers will wish to treat the total group of items on a questionnaire as a single measuring scale, and from them obtain a single overall score on whatever they are interested in measuring. This is a common practice, especially with the measurement of attitudes. A typical attitude scale is composed of a number of questions/statements selected and put together from a much larger number of questions/statements according to certain statistical procedures. Some of these procedures, called scaling techniques, are discussed in this chapter.

A distinction is needed, however, between two ways in which the term scale is used in this manual. An attitude scale could be constituted of items each one of which employs a response scale. Aspects of response scales are discussed in Chapter VII on "Response Anchoring." A component of score could be achieved on each item. Adding these item scores together - which means considering the whole set of items as a scale - produces a total attitude score for the individual respondent.

There are, generally speaking, two general methods for the construction of scales such as attitude scales. The first method makes use of a judging group and one of the psychological scaling methods developed by Thurstone, as discussed in Section V-B. It results in a set of statements being assigned scale values on a psychological continuum. The continuum may be favorableness-unfavorableness, like-dislike, or any other judgment. The psychological scaling methods, therefore, have considerably greater application than for the scaling of attitudes. They can be used to scale statements or objects. They have been used, for example, to determine the perceived favorableness of words and phrases commonly used as rating scale response alternatives, as discussed in Chapter VIII.

The second general method is based on the direct responses of agreement or disagreement with attitude statements and does not result in a set of statements being assigned scale values on a psychological continuum. Both the Likert and Guttman scales discussed in Sections V-C and V-D are examples of this latter method.

For information (relating to attitude scaling and scaling techniques) beyond that contained in this manual, the following references may be consulted.

1. Babbitt, B. A., & Nystrom, C. O. (1985). Training and human factors research on military systems. Questionnaires: Literature survey and bibliography. Fort Hood, TX: Army Research Institute for the Behavioral and Social Sciences.

2. Church, F. (1983, June). Questionnaire construction manual for operational tests and evaluation. Prepared for the Deputy Commander of Tactics and Test, 57th Fighter Weapons Wing/DT, Tactical Fighter Weapons Center (TFWC), Nellis AFB, NV.

3. Edwards, A. L. (1957). Techniques of attitude scale construction. New York: Appleton-Century-Crofts.

4. Eggemeier, F. T., Crabtree, M. S., & La Point, P. A. (1983, October). The effect of delayed report on subjective ratings of mental workload. Proceedings of the Human Factors Society 27th Annual Meeting, 139-143.

5. Eggemeier, F. T., Crabtree, M. S., Zingg, J. J., Reid, G. B., & Shingledecker, C. A. (1982). Subjective workload assessment in a memory update task. Proceedings of the Human Factors Society 26th Annual Meeting, 643-647.

6. Eggemeier, F. T., McGhee, J. Z., & Reid, G. B. (1983, May). The effects of variations in task loading on subjective workload rating scales. Proceedings of the IEEE 1983 National Aerospace and Electronics Conference, Dayton, OH, 1099-1105.

7. Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.

8. Gulliksen, H., & Messick, S. (Eds.) (1969). Psychological scaling: Theory and applications. New York: John Wiley.

9. Lemon, N. (1974). Attitudes and their measurement. New York: John Wiley.

10. McIver, J. P., & Carmines, E. G. (1981). Unidimensional scaling. Sage University Paper series on quantitative applications in the social sciences, 07-024. Beverly Hills and London: Sage Publishers.

11. Moroney, W. F. (1984). The use of checklists and questionnaires during system and equipment test and evaluation. Shrivenham, England: NATO Defense Research Group Panel VIII Workshop, Applications of Systems Ergonomics to Weapon System Development, Royal Military College of Science, Vol 1, C-59-C-68.

12. Nunnally, J. C. (1967). Psychometric theory. New York: McGraw-Hill.

13. Thurstone, L. L. (1959). The measurement of values. Chicago: University of Chicago Press.

14. Torgerson, W. S. (1958). Theory and methods of scaling. New York: John Wiley.

B. Thurstone Scales

This section discusses three scaling methods developed by L. L. Thurstone. Thurstone investigated rank order scales and how to compare psychological variables. He developed the law of comparative judgment with an underlying assumption that the degree to which any two stimuli can be discriminated is a direct function of the difference in their status as regards the attribute in question. Thurstone generated three new scaling methods based on his law of comparative judgment. The three scaling methods are known as equal appearing intervals, paired-comparison, and successive intervals. For additional detail, see the texts referred to in Section V-A.

1. Method of Equal Appearing Intervals

Thurstone's method of equal appearing intervals assumes that a group of statements of opinion about a particular issue could be ordered on a continuum of favorableness-unfavorableness, and that the ordering could be such that there appears to be an equal distance between the adjacent statements on the continuum.

The following steps are followed in the method of equal appearing intervals:

a. From the literature or pilot interviews, a large number of statements (100 to 200) are compiled about the attribute or object of an attitude under study. Irrelevant, ambiguous, or poorly worded statements would not be selected.

b. A number of judges, at least 50, are obtained. They should be similar to those individuals who will respond to the final statements on the questionnaire. The judges independently sort each statement into one of 11 piles. The first pile is defined as "Unfavorable" or "Most unfavorable," the middle or sixth pile is defined as "Neutral," and the eleventh pile is defined as "Favorable" or "Most favorable." The other piles are left undefined. The judges are told that the intervals between piles or categories are to be regarded as subjectively equal. They are also instructed to ignore their own agreement or disagreement with each item, and to judge each item in terms of its degree of favorableness-unfavorableness.

c. The scale value for each item is usually determined by computing its mean or median, over all judges.

d. Twenty to 25 statements with little dispersion in their scale values are then selected for use. The statements are selected so that the intervals between statements' scale values are approximately equal and/or are relatively equally spaced on the psychological continuum.

e. The finally selected statements are usually placed in random order for presentation to respondents. The respondents are asked to indicate which statements they agree with, and which they disagree with.

f. The respondent's score is the mean or median scale value of those statements for which he/she marked "Agree."

Some considerations for use of the Equal Appearing Intervals method are:

a. The method of equal appearing intervals is designed to provide an interval scale as its output. The scale is at least ordinal (ranked).

b. The method is useful when there are a large number of statements involved.

c. Scale values from widely differing groups of judges appear to correlate highly with one another so long as judges with extreme views are eliminated.

d. Graphic or numerical rating scales can be used by the judges instead of having the statements sorted into piles. Though 11 categories are usually used, some other number can be employed.

e. There have been some psychometric questions about the unidimensionality of Thurstone scales. Even though research has been mixed as to which scaling methods are best, there is some evidence that Likert and Guttman scales may be sounder. Actual scale format does not seem to be as important as the actual developmental procedures in the construction of the scale.

2. The Method of Paired Comparisons

Thurstone developed a procedure for deriving an interval scale based upon what has been called the Law of Comparative Judgment. Basically, it is a method by which statements such as "A is stronger than B," "B is stronger than C," etc., are used to provide a scale with interval properties. The objects or statements to be ranked are presented two at a time, and the respondent is asked to choose between them. All possible combinations of pairs have to be presented. Hence the procedure becomes very cumbersome when there are more than 15 or so items. The determination of scale values is also laborious. Since the procedure is not used much in applied research, additional detail is not presented here.

3. The Method of Successive Intervals

The method of successive intervals is similar to the method of equal appearing intervals. However, no assumption is made concerning the psychological equality of the category intervals.

It is only assumed that the categories are in correct rank order and that their boundary lines are relatively stable. The procedure involves estimating the widths of the categories along the psychological continuum. From these reference points, the scale values of the statements can be obtained. Research has shown that there is a linear relationship between scales constructed by the method of paired-comparisons and by the method of successive intervals.

4. New Applications for Thurstone Scales

When Thurstone developed the law of comparative judgment, his scaling techniques were considered a major advancement. Thurstone scales continue to be used in survey research, although other scaling methods have gained popularity, such as Likert and Guttman scales. There have been instances when rank order scaling procedures have been integrated into other complex systems. An illustration of this is the delta scalar method used by the U.S. Navy and the Air Force Aerospace Medical Research Laboratory. The delta scalar method is a complex system of rank ordering found in the Mission Operability Assessment Technique and Systems Operability Measurement Algorithm (U.S. Navy, and the Subjective Workload Assessment Technique (U.S. Air Force). These systems involve establishing a rank order scale that is converted to an interval scale. More research will be required to determine how functional, reliable, and valid these new procedures will be. The procedures for embedding rank order methods into other scales is complicated and beyond the scope of this manual.

C. Likert Scales

The Likert method of scale construction was developed because the Thurstone procedures require extensive work and make assumptions regarding the independence of item statements. The Likert method assumes that all statements reflect the same attitude dimension and are hence related to each other. The Likert approach does not assume equal intervals between the scale values. It is sometimes called the method of summated ratings.

The steps in Likert scale construction are as follows:

1. Item Construction

Design an initial set of items to measure an attribute. Statements are classified in advance as "Favorable" or "Unfavorable." No attempt is made to find an equal distribution of statements over the whole range of the attitude of concern, and no attempt is made to scale the statements.

2. Item Selection

Likert proposed the use of correlation analyses and analyses based on the criterion of internal consistency to evaluate the ability of individual items to measure an attribute.

a. A pretest is conducted. In the pretest, the respondents indicate their degree of agreement with every statement, usually using five response alternatives: strongly agree, agree, undecided, disagree, and strongly disagree. Each descriptor is assigned a numerical weight (e.g., 4, 3, 2, 1, 0) usually based on a given series of integers in arithmetical sequence. Each respondent is assigned a score that represents the summation of weights associated with each item checked.

b. Criterion of internal consistency compares the difference between mean responses to an individual item compared to high and low subgroups. Subgroups consist of 25% of the respondents at each extreme of the scale.

c. The criterion of internal consistency includes differences in subgroup size and different distributions of responses between subgroups.

d. The t test provides an accurate indication of the degree to which an item differentiates between high and low subgroups.

$$t = (\bar{X}_H - \bar{X}_L) / \sqrt{(S_H^2/n_H) + (S_L^2/n_L)}$$

$\bar{X}$ = mean item response of subgroup

S^2 = item variance of subgroup

n = size of subgroup

e. The criterion of internal consistency analysis and the correlation analysis may lead to different conclusions regarding the selection of items. It is recommended that both types of item analyses be used to assist in determining which items to retain.

f. Correlational analysis focuses on how strongly the item is related to the total scale score.

$$r_{i(T-i)} = (r_{iT}\,\sigma_T - \sigma_i) / \sqrt{(\sigma_T^2 + \sigma_i^2) - 2r_{iT}\,\sigma_T\sigma_i}$$

r_{iT} = correlation between item and total score

σ_T = standard deviation of the total score

σ_i = standard deviation of the item score

The greater the number of items, the less each item will contribute to the variance of the scale. Each item will contribute more bias for scales that have only a few items.

g. Each item is treated as a predictor of the respondent's total score. Items with low item-to-total correlations should be eliminated from the scale. Items that do not discriminate between groups with extreme attitudes (25% of the respondents at each extreme of the scale) should be eliminated. This procedure leaves us with the items that will comprise the final score.

3. Item Scoring

a. Calculate scale scores by summing the response scores for each item given the following values. Favorable statements receive a value of 4 for "Strongly agree" and a value of 3 for "Agree." The midpoint response alternative "Undecided" receives a value of 2. Unfavorable statements receive a value of 1 for "Disagree" and a value of 0 for "Strongly disagree." High scores always indicate a favorable attitude, and low scores always indicate an unfavorable attitude.

b. Interpretation of individual scoring is defined relative to the group. Each of the individual attitude scores is expressed as a deviation from the mean of the group. The score of any individual relative to the mean of the group is:

$$X - \bar{X}$$

X = individual score

$\bar{X}$ = group mean

The scores are converted into Z scores by dividing each individual score by the standard deviation of the sample. A Z score will identify the position of the respondent's score in relation to the mean of the distribution. Using the curve as a distribution of observations, the Z score can describe the location of the score along the horizontal axis. A Z score distribution maintains the same shape as the set of raw scores from which it was derived.

$$Z = \frac{X - \bar{X}}{S}$$

Z scores indicate how many standard deviations the score lies above or below the mean. The mean is always zero, and the standard deviation of any set of Z scores is always 1. Z scores can be used to compare scores from different distributions so long as the distributions have approximately the same shape.

4. Reliability of the Summated Scale

To compute the reliability of the Likert scale, the coefficient alpha is recommended.

$$\alpha = N\bar{\beta} / [1+\bar{\beta}(N-1)]$$

N = number of items

$\bar{\beta}$ = mean interitem correlation

The alpha coefficient provides an estimate of reliability based on the interitem correlation matrix.

Factors to be taken into consideration when deciding whether to use Likert scales include:

1. Likert scales take less time to construct than Thurstone scales. They are one of the most widely used scales for attitude surveys.

2. It is possible to construct scales by the Likert and Thurstone methods which will yield comparable scores.

3. Likert scales have only ordinal properties. If there is a large dispersion about a respondent's mean score, however, even those properties have limited meaning. If the sole purpose of a scaling procedure is to rank respondents according to the degree to which they hold some attitude, then Likert scales are efficient because of their ease of administration.

4. In addition to lacking metric properties, Likert summated scores lack a neutral point. The interpretation of a score cannot be made independently of the distribution of scores of some defined group. Only the summation of the items measure the attitude. Percentile or deviation-type norms can be calculated if the sample size is large enough.

5. For the same number of items, scores from Likert scales may be more reliable than scores from Thurstone scales.

6. Likert and Guttman scales both appear to be superior to Thurstone scales.

D. Guttman Scales

Guttman scaling was developed as an alternative to Thurstone and Likert methods of attitude scaling. Guttman's apprcach to scaling is known as scalogram or scale analysis. It is a deterministic model; it considers its scales are close to being rulers-measures of length. The essence of the method is to determine whether a series of statements can be appropriately scaled. An attempt is made to identify a set of statements which actually reflect a unidimensional scale and have a cumulative nature. When the goal is achieved, two or more persons receiving the same score will have responded in the same way to all of the statements.

As an example, the following four questions comprise a Guttman scale:

	Yes	No
a. The United Nations is mankind's savior	___	___
b. The United Nations is our best hope for peace	___	___
c. The United Nations is a constructive force in the world	___	___
d. We should continue our participation in the United Nations	___	___

The expected pattern of responses to these questions is "triangular."

	Person				
Item	1	2	3	4	Scale Score
a	x				1
b	x	x			2
c	x	x	x		3
d	x	x	x	x	4

This means that, for persons who answers yes to item "a," there is a high probability that they will answer yes to the other items. A person who says no to "a" but yes to "b" has a high probability of answering yes to the other items, and so on. The model anticipates that the perfect relationship between the scale score and the item score will be violated. The degree of deviation that is acceptable is established by criteria, and measured by a coefficient of reproducibility.

Guttman scaling is considered psychometrically more robust than Likert or Thurstone scaling. The coefficient of reproducibility (CR) could be used to evaluate the degree of scalability of empirical data. The Guttman model calls for assigning scale scores only when the coefficient of reproducibility (CR) is greater than .90. The formula is as follows:

$$CR = 1.0 - (\#\ errors)/total\ responses$$
$$= 1.0 - (\#\ errors)/[(\#\ items) \times (\#\ respondents)]$$

For example, a respondent who rates three items positively out of n items composing the Guttman scale would be considered to have responded to three specific items which would be considered the three items most acceptable to the population of respondents. The interpretation of a response to three items on a Likert scale would be that the respondent had rated favorably any three items of n stimuli.

The major steps in scalogram analysis are too complex to summarize here, but are found in some of the references in Section V-A. Procedures are available for:

1. Measuring the amount of error due to imperfect scalability.

2. Ordering the statements so that the response patterns provide the least amount of error.

3. Determining the extent to which the data approximate the perfect case.

4. Improving the scalability of the statements via category combinations, statement discarding, etc.

There have been many critics of scalogram analysis. Some feel that there is no really effective way of selecting good items by this approach. However, the procedure is considered useful if one is concerned with unidimensionality or if one wishes to examine small changes in attitudes. Guttman scaling is primarily used in the construction of attitude surveys as well as in the construction of mixed standard scales. It may be possible to construct other mixed standard scales for surveys that measure other factors in addition to job performance. It is laborious to construct Guttman scales. No instances of past use in field testing situations are known.

Even though Guttman's approach to scale analysis has not been used in field testing situations, it is being used by the armed services for other applications.

Adaptive testing is based on a Guttman method of scaling and adaptive testing is being investigated by the armed services. The Armed Services Vocational Aptitude Battery is being developed for computer-adaptive testing by the Navy Personnel Research and Development Center. Each time a question is asked, there is a recalculation of probabilities so that the next item selected is based on the subject's response to the previous item. This allows for estimating the respondent's future performance level as a way to select the next item. The items are administered on a computer, and each respondent receives a different set of questions.

Adaptive testing requires a large sample for its development. It has been primarily used as an ability test with multiple choice questions. There have been other types of applications such as for interviewing. The armed forces are a leader in adaptive testing. Even so, currently, this model does not appear to be viable for OT&E because of the large samples, and the lead time required for development.

E. Other Scaling Techniques

Numerous other scaling techniques and combinations of methods are reported in the literature. A discussion of them, however, is outside the current scope of this manual.

Chapter VI: Preparation of Questionnaire Items

A. Overview

Once a decision has been made regarding the type or types of items that are to be used in a questionnaire (see Chapter IV), attention must be given to the actual development of the items. This chapter addresses the following development topics: mode of questionnaire items; wording of items for both question stems and response alternatives; difficulty of items; length of question stem; order of question stem; number of response alternatives; and order of response alternatives. The related topic of response anchoring is considered in Chapter VII.

As used in this manual, a distinction has been made between a questionnaire item, a question stem, and response alternatives. A questionnaire item has both a question stem and response alternatives. The response alternatives are the answer choices for the question. (They are sometimes called "options.") The question stem is that part of the item that comes before the response alternatives.

B. Mode of Items

Questionnaire items are usually presented to a respondent in printed form. However, it is possible to present items or stimuli pictorially. There is some evidence that there are no significant differences in subjects' responses to verbal and pictorial formats. The evidence is conflicting, since anchoring endpoints with pictorial anchors for bipolar scales has proven difficult in establishing meaning. Researchers were not able to verify that the pictorial anchors were actually antonyms. This could affect the bipolar assumptions of the scales. Using a pictorial format may facilitate obtaining responses from respondents with limited verbal comprehension, who might have difficulty responding to questions employing lengthy definitions of concepts or objects. If pictures are used, they should be pretested for clarity of their presentation of the concept or object to be evaluated.

For group administration of a questionnaire with pictorial anchors, it would be possible to use color slides and rating forms with replicas of the slides. In cases where it is known that the respondents have very low reading ability, it may be desirable to present the questionnaire orally. A tape player-recorder may be used for this purpose also.

C. Wording of Items

The wording of questionnaire items is a critical consideration in obtaining valid, relevant, and reliable responses. Consider, for example, the following three questions that were administered by Payne (see reference below) to three matched groups of respondents:

a. "Do you think anything should be done to make it easier for people to pay doctor or hospital bills?"

b. "Do you think anything could be done to make it easier for people to pay doctor or hospital bills?"

c. "Do you think anything might be done to make it easier for people to pay doctor or hospital bills?"

These questions differed only in the use of the words "should," "could," or "might," terms that are often used as synonyms even though they have different connotations. The percent of "Yes" replies to the questions were 82, 77, and 63, respectively. The difference of 19% between the extremes is probably enough to alter the conclusions of most studies.

A number of matters related to the wording of questionnaire items are considered in this section. Some of the suggestions made are based upon experimental research. Others are based upon experience, intuition, and commonsense. Several sources offering principles of question wording are:

a. Roslow, S., & Blankenship, A. B. (1939). Phrasing the question in consumer research. Journal of Applied Psychology, 23, 612-622.

b. Jenkins, J. G. (1941). Characteristics of the question as determinants of dependability. Journal of Consulting Psychology, 5, 164-169.

c. Blankenship, A. B. (1942). Psychological difficulties in measuring consumer preferences. Journal of Marketing, 6, 66-75.

d. Payne, S. L. (1963). The art of asking questions (Rev. ed.). Princeton, NJ: Princeton University Press.

e. Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, Inc.

1. Formulation of the Question or Question Stem

a. General comments regarding items and question stems. Issues that should be noted concerning the general structure of questions and question stems are:

(1) Question stems may be in the form of an incomplete statement, where the statement is completed by one of the response alternatives, or in the form of a complete question. See Figure VI-C-1 for examples.

Figure VI-C-1

Example of Question Form (Item 1) and
Incomplete Statement Form (Item 2) of Stem

1. How qualified or unqualified for their jobs are most Army NCOs? (Check one.)

_____ Very well qualified

_____ Qualified

_____ Borderline

_____ Unqualified

_____ Very unqualified

2. Check one of the following. Most Army NCOs are:

_____ Very well qualifed for their jobs.

_____ Qualified for their jobs.

_____ Borderline.

_____ Unqualified for their jobs.

_____ Very unqualified for their jobs.

The choice between these two methods should depend on which of the two permits simpler and more direct wording for the item in question. Not all of the items in a questionnaire need to be in the same form.

(2) All questionnaire items should be grammatically correct.

(3) All stems should be as neutrally expressed as possible, and the respondents should be permitted to indicate/select the direction of their preference. If this is not done, the stems may influence the response distribution. If the stems cannot be expressed neutrally, then alternate forms of the questionnaire should be used.

(4) Respondents may not answer an item if they are not able to give the information requested. Therefore, care should be exercised in the wording of the question, so that it does not call for information not possessed by the respondents. If the respondent is not able to answer the item, the response option should permit the respondent to say he/she "doesn't know." The questionnaire designer should have determined during pretesting whether a "Don't Know" response option should be included.

b. Accuracy and completeness of question stems.

(1) The stem of an item should be accurate, even though inaccuracies may not influence the selection of the response alternative.

(2) The question stem, in conjunction with each response alternative, should present the question as fully as necessary to allow the respondent to answer. It should not be necessary for the respondent to infer essential points. An example of an insufficiently informative question stem is given as item 1 in Figure VI-C-2. It is insufficient in that no specification is given as to who should carry the scopes. (The response alternatives are also insufficient since the respondent is not allowed to say "None.") Two or three questions might be needed to obtain all the information desired. Item 2 in Figure VI-C-2 is one revision that makes the question stem sufficient.

(3) Generally, materials which are common to all response alternatives should be contained in the stem, if this can be done without the need for awkward wording.

(4) In forming questions which depend on respondents' memory or recall capabilities, the time period a question covers must be carefully defined. The "when" should be specifically provided.

Figure VI-C-2

An Insufficiently Detailed Question Stem, Plus Revision

1. How many starlight scopes should be issued to a rifle squad?

   _____ 1

   _____ 2

   _____ 3

   _____ 4

   _____ 5

2. Place a check in front of each squad member's "name" below that you believe should be issued a starlight scope:

_____ Squad Leader	_____ Fire Team 2 Leader
_____ Fire Team 1 Leader	_____ Automatic Riflemam
_____ Automatic Rifleman	_____ Grenadier
_____ Grenadier	_____ Rifleman
_____ Rifleman	_____ Rifleman

(5) Question stems and response alternatives should be worded so that it is clear what the respondent meant. Consider the question "Should this cap be adopted, or its alternate?" If the respondent answers "Yes," it would still be unclear which cap ("this cap" or "its alternate") should be adopted.

c. Positive versus negative wording.

(1) Alternative wording can produce demonstrable effects on survey results.

(2) There may be a tendency for the direction of the question stem to be chosen in the response alternative.

(3) Studies have indicated that it is usually undesirable to include negatives in question stems (unless an alternate form with positives is also used for half of the respondents).

(4) Questions worded in positive terms are preferred by respondents to questions in negative terms (if alternate forms are not being used). Questions worded negatively may be confusing, or n ative words may be overlooked.

(5) If it seems necessary to have a particular question in negative form, the negative word (e.g., not, never) should be underlined or italicized. Care should also be taken that there are no double negatives, as they are frequently misinterpreted.

(6) A question worded in negative terms can often be improved by rephrasing it in positive terms.

(7) There is evidence to indicate that positively-worded items may in some instances receive higher mean responses than negatively-worded items. However, these findings were not statistically significant. There are conflicting research results where positive and negative wording of items did not affect the responses.

d. Definite versus indefinite article wording. The indefinite articles, "a" or "an," would be used in a question such as "Did you see a demonstration of the new night vision device?" A comparable question using the definite article "the" would be, "Did you see the demonstration of the new night vision device?" There is some evidence that changing from "a" to "the" reduces the level of suggestibility of an item. However, there is not enough evidence to warrant a firm conclusion.

e. First, second, and third person wording. An example of a statement written in the first person is, "Army NCOs are understanding of my needs and problems." A statement in the second person is, "Army NCOs are understanding of your needs and problems," while one in the third person is, "Army NCOs are understanding of the needs and problems of their men." It is preferable that the framework of questions be consistent for all questions in a questionnaire, so that responses are comparable. A respondent's opinion of events affecting his/her own person is ofte quite different than his/her opinions of the effects of the same events on others. Hence, questions written in the first or second person may elicit entirely different responses than the "same" question written in the third person.

There are occasions where each person (first, second, or third) is appropriate. For example, the third person should probably be used when it is desired to elicit information that might be considered too personal for a person to answer about himself/herself. The third person may also be used in attempts to elicit information about the feelings inherent in a minority of respondents, but about which many more respondents may be

aware, such as in the statement, "The Army is ahead of most areas of civilian life in reducing racial discrimination." In other cases, the first or second person form is not applicable, such as in "The Army is essential for the defense of the country." Also, the use of the third person permits a far larger number of personnel to answer the questions, since some first person questions that are inapplicable to many individuals become applicable when in the third person. Instances may occur where respondents are asked a question twice, once to discover how they personally feel about the issue (using first or second person), and then to discover what they judge others' feelings on that issue are (using the third person). Some personally-worded items may be perceived as more specific to the experience of the respondent. This may possibly provide results that have greater accuracy for items that are non-threatening. Generally, however, the use of the third person appears preferable.

f. Loaded and leading questions. Loaded and leading questions should be avoided. Although the questionnaire writers may not deliberately attempt to distort the distribution of responses, they may sometimes do so unintentionally.

In Figure VI-C-3, item 1 should be revised to maintain neutrality by removing the adjectives applied to the rifles. It is true that the M16 weighs less and fires more rounds faster, but there are other characteristics (accuracy, lethality given a hit, etc.) that are not cited. Hence, the question is loaded because it only presents some of the data relevant to comparing the rifles.

Items 2 and 3 in Figure VI-C-3 show loading of a different type. In item 2, analysis of the available alternatives leaves the impression that the writer of the question thinks at least some should not have a full automatic selector. Analysis of the alternatives in item 3 leads to the suspicion that the writer of the question believes there should be at least one grenade launcher in the rifle squad, since a response alternative of zero grenade launchers was not provided.

There are many additional ways that questions can be loaded. One way is to provide the respondent with a reason for selecting one of the alternatives, as with the question, "Should we increase taxes in order to get better schools, or should we keep them about the same?" A question can also be loaded by referring to some prestigious individual or group, as in, "A group of experts has suggested...Do you approve of this, or do you disapprove?"

Figure VI-C-3

Examples of Loaded Questions

1. Which rifle do you prefer, the lighter, faster shooting M16 or the heavier, slower firing M14?

   _____ M16

   _____ M14

2. Should every rifleman in the rifle squad have a full automatic selector on his rifle?

   Yes_____

   No_____

   If no, how many should?_____

3. How many grenade launchers (M79) do you desire in the rifle squad?

   _____ 1

   _____ 2

   _____ 3

   _____ 4 or more

Leading questions are similar to loaded questions. Two examples are shown in Figure VI-C-4. The problem is that most people are reasonably cooperative and like to help. If they can figure out what is wanted, they will often try to comply. The items in Figure VI-C-4 were actually used in the collection of data in a field test. As might be expected, the impression received from an analysis of the results is that men are, in general, highly motivated, and use good noise discipline during movement. (These items also allow respondents to avoid criticizing, and to give socially desirable answers.)

Figure VI-C-4

Examples of Leading Questions

1. Do you think your men were pretty highly motivated on this exercise?

   Yes_____

   No_____

2. Were they pretty good at using good noise discipline during movement?

   Yes_____

   No_____

The best way to avoid loaded questions is to find a devil's advocate to review them or to pretest the items on someone who holds opposite or minority views. Another check is to ask yourself what you think, what someone who disagrees with you would think, and whether your response alternatives would give the respondents a chance to present their views.

Not every change in wording will have a significant effect on the item. This provides a measure of latitude in the design of the items. Blatant attempts to bias an item by tone of wording are not so likely to succeed. Research indicates that blatant language may have no effect on responses. There is no convincing evidence that respondents with strong attitudes toward a topic would be less influenced by the tone of wording than respondents who did not have a strong attitude toward the topic.

There are times when loaded questions probably should be used. This is when, without loading, the question would pose an ego-threat to the respondents, so that they might give an untruthful reply. The loading removes the ego-threat so that a more valid response can be obtained. An example might be, "Many people are not able to get as much schooling as they would like. What was the last grade you completed in school?"

g. Embarrassing or self-incriminating questions. Respondents should not be asked embarrassing or self-incriminating questions. Consider the question, "Did you clean your weapon regularly in Vietnam?" It is asking respondents who did not clean their rifles regularly to expose themselves to possible embarrassment. Thus, one would expect the percentage of "No" responses to fall short of the true percentage not cleaning their weapons "regularly."

Occasionally questionnaires cover topic areas that are sensitive, and may be perceived as threatening by the respondents. For this type of questionnaire, threatening questions elicit greater under-reporting when closed-end questions are used. Thus, open-ended questions are appropriate for threatening topics. Longer questions, using the language of the respondent, seems to decrease unwanted response effects for threatening questions. In addition, willingness to answer threatening questions is increased by assuring respondents that their answers will be treated confidentially. An example of a threatening question is presented in Figure VI-C-5 which illustrates a longer, open-ended item used with threatening content.

Figure VI-C-5

Example of a Threatening Question

Please describe and explain in your own words any problem in your unit that might be caused by the use of too much alcohol, marijuana, or hard drugs by upper-ranking officers, senior NCOs, or supervisors.

_____

_____

_____

_____

_____

h. Questions that ask respondents to go against basic inclinations.

Many people are reluctant to criticize, though they enjoy giving praise. Thus, a question that allows respondents to avoid criticism will bias their answers; similarly, a question that offers them the opportunity to criticize may bias responses because they will not wish to do so. Figure VI-C-6 illustrates this.

Figure VI-C-6

Example of a Question
Asking the Respondent to Criticize

1. Was your unit's use of fire and maneuver correct, and in accordance with current Army doctrine?

   Yes_____

   No_____

   If no, why not?_____

   _____

The question in Figure VI-C-6 asks the respondents either to criticize their unit or to avoid criticism. Some respondents might answer "No" if they have an important point to make. However, a substantial number of others will wash their hands of the whole affair and answer "Yes," although they might feel that performance was not completely correct.

i. Inclusion of different subjects into the same question. Compound questions should be avoided. These are questions that require a respondent to give the same assessment of two or more issues/characteristics or aspects of the subject. Respondents must be allowed to make separate assessments of each issue. Consider, for example, item 1 in Figure VI-C-7. Most respondents would probably want to rate completeness and accuracy differently, since in most situations research has shown that they are negatively correlated. Therefore, the two aspects of performance should be rated separately, as shown in items 2 and 3 of Figure VI-C-7.

Figure VI-C-7

Examples of Compound Questions and Alternatives

1. How complete and accurate was the surveillance information?

_____ Very satisfactory

_____ Satisfactory

_____ Borderline

_____ Unsatisfactory

_____ Very unsatisfactory

2. How complete or incomplete was the surveillance information?

_____ Very complete

_____ Fairly complete

_____ Borderline

_____ Fairly incomplete

_____ Very incomplete

3. How accurate or inaccurate was the surveillance information?

_____ Very accurate

_____ Fairly accurate

_____ Borderline

_____ Fairly inaccurate

_____ Very inaccurate

It may be noted that in item 2 of Figure VI-C-7 both "complete" and "incomplete" are included. Similarly, both "accurate" and "inaccurate" are in the stem of item 3. To use only one (e.g., "complete") in the stem would tend to inflate the number of respondents selecting that alternative.

j. Use of giveaway words. Avoid words which lead the careful thinker to respond in the negative, while others, thinking less carefully, respond in the positive. Consider for example the question, "Do you feel that your unit did its best in all contacts over the past six months?" One wonders if any unit can do its actual best, except very rarely. The word "all" makes this an even more difficult question to answer positively.

k. Ambiguous questions. Vague or ambiguous words or questions should be avoided. For example, the question "What is your income?" is not sufficiently specific. The respondents may give monthly or annual income, income before or after taxes, their income or the family income, etc.

As another example, consider item 1 in Figure VI-C-8.

Figure VI-C-8

Example of Ambiguous Question and Alternative

1. Did you clean your rifle regularly in Vietnam?

_____ Yes

_____ No

2. How often, on the average, did you clean your rifle in Vietnam?

_____ Every day

_____ Once every two days

_____ Once every three days

_____ Once every four days

_____ Other (please specify): _____

_____

Use of the word "regularly" without specification of the time interval between cleanings is a defect in the question. A respondent could justify a "yes" by thinking to himself/herself: "Sure, I cleaned it regularly - once every four months!" Because of the self-exposure involved, the questionnaire item approach to this topic is probably not capable of providing an accurate estimate, but rewording could still make the amount of underestimation less. So, if the data cannot be collected by field inspection, the revised questionnaire item could read like item 2 in Figure VI-C-8.

Items are sometimes loaded because the wording is ambiguous, coerces agreement, or uses jargon or technical words that are not understandable. Review of items for illogical response patterns may be useful when respondents have less education. Figure VI-C-9 illustrates items which were highly ambiguous in their wording. Some respondents did not consider the first item in a literal sense for its impact on subsequent items. This set of items obtained many illogical response patterns.

Figure VI-C-9

Example of Ambiguity of Wording

"Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?"	YES, NO, NOT SURE
"Would you approve if the citizen . . ."	
A. "had said vulgar and obscene things to a policeman?"	YES, NO, NOT SURE
B. "was being questioned as a suspect in a murder case?"	YES, NO, NOT SURE
C. "was attempting to escape from custody?"	YES, NO, NOT SURE
D. "was attacking the policeman with his fists?"	YES, NO, NOT SURE

When items are long and negatively worded, they may create ambiguity. This ambiguity seems to result in an increased number of responses in the middle alternative. Pretesting the items would provide the opportunity for modification of items by obtaining feedback from the respondents on issues related to the complex meaning of any technical words, and any multiple meanings of words. Use the language of the respondents in developing and refining items.

2. Formulation of the Response Alternatives

When formulating the response alternatives portion of a questionnaire item, the following points should be kept in mind:

a. All response alternatives should follow the stem both grammatically and logically, and, if possible, be parallel in structure.

b. If it is not known whether or not all respondents have the background or experience necessary to answer an item (or if it is known that some do not), a "Don't know" response alternative should be included.

c. When preference questions are being asked (such as "Which do you prefer, the M16 or the M14 rifle?"), the "No preference" response alternative should usually be included. The identification of "No preference" responses permits computation of whether or not an actual majority of the total samples are pro or con.

d. Respondents with a low educational level have a propensity to use the "Don't know" response alternative.

e. When the "Don't know" response alternative is used, it should be set apart from other responses to avoid confusing it with the endpoint or the midpoint of the rating scale.

f. Content items can be developed which will indicate whether a subject has knowledge regarding the topic in question. If the subject has little topic knowledge, and there is not a "Don't know" category, there is the potential for greater rating error.

g. The use of the "None of the above" option or variants of it, such as "Not enough information," is sometimes useful.

h. The option "All of the above" may on rare occasions be useful. It seems more appropriate to academic test questions than to the questioning of field test participants.

i. For most items, the questionnaire writer desires the respondent to check only one response alternative. Use of the parenthetic "(Check one.)" should eliminate the selection of more than one alternative. It is very important to make it clear to the respondents that they may check more than one alternative in those fairly rare instances where the questionnaire writer does wish to permit this.

j. In some instances, response categories as long as a sentence may be more desirable than short descriptors. In rare cases, numbers may be used without verbal descriptors, if the numbers have been previously defined. It does not seem to matter if the response alternatives are numerical, verbal (one word), or phrases. No one type of response alternative has proven superior to another.

k. When the quality of the item is high and the data is available, response alternatives can be selected which have standard deviations less than 1.00 (see Section VIII-E).

l. There is some evidence that responses to scales labeled at only the extreme ends have been skewed toward the positive end of the scale. Fully labeled scale points may encourage a more balanced response distribution.

m. Number of response alternatives is discussed in Section VI-G, order of response alternatives in Section VI-H, response anchoring in Chapter VII, and the order of perceived favorableness of commonly used words and phrases in Chapter VIII.

3. Expressing Directionality and Intensity in Stem Versus Response Alternatives

In item 1 of Figure VI-C-10, directionality (in this case, satisfaction) is expressed in the question stem.

Figure VI-C-10

Alternate Ways of Expressing Directionality and Intensity

1. The M16 is a satisfactory rifle.

_____ Agree

_____ Disagree

2. The M16 is

_____ a satisfactory rifle.

_____ an unsatisfactory rifle.

3. The behavior of civilian employees of the PX toward enlisted personnel is extremely offensive.

_____ Agree

_____ Disagree

4. The behavior of civilian employees of the PX toward enlisted personnel is

_____ very offensive.

_____ somewhat offensive.

_____ neutral.

_____ somewhat pleasant.

_____ very pleasant.

In item 2, the directionality is expressed in the response alternatives. In item 3, the stem contains terms of intensity and directionality, while these terms are located in the response alternatives in item 4. Item 2 is preferred to item 1, and item 4 is strongly preferred to the item 3 approach. The rationale for this preference is similar to the discussion of positive versus negative terms. Those who check "Disagree" to item 3 have not been permitted to indicate what it is they would agree with, (e.g., those who feel employees are offensive but not extremely offensive would have to check "Disagree," as would those who feel employees are very pleasant), whereas the construction of item 4 does permit them to do so. It would take five versions of item 3 to correct this deficiency and achieve the coverage of opinion incorporated by the response alternatives of item 4.

D. Difficulty of Items

1. One of the major recommendations advanced by almost every general source on how to write sound questionnaires is "keep it simple." Logic dictates that words used in surveys should not have multiple meanings, nor should they be beyond the level of vocabulary of the typical respondent. Words, phrases, and sentence structures that the respondent can understand should be used.

   Consider item 1 in Figure VI-D-1. It contains too many hard to understand words. Many respondents would have difficulty understanding either the question or the response alternatives. In the revision in item 2, the words have been simplified and a "catch-all" open-ended response alternative added (to catch all other reasons).

Figure VI-D-1

Example of Hard to Understand Item and Alternative

1. In the highly specialized counterinsurgency environment represented by the basically internecine affair in Vietnam, what would you say should represent the basic essence of our rationale for continuation of our involvement?

   _____ Prolongation of attrition of enemy forces, in order to reduce the level of threat to South Vietnam.

   _____ Orderly transfer of military responsibility to the host country, in order to produce stabilized competency to deal with any future internal distrubances.

2. What is our main reason for staying in Vietnam? (Check one)

   _____ To reduce the threat to South Vietnam by continuing the destruction of enemy forces.

   _____ To assure South Vietnam's survival while it takes over responsibility for its own protection.

   _____ Other (specify) _____

   _____

It should not be assumed that the respondent will understand what the question writer is talking about. Consider, for example, the question "Which do you prefer, dichotomous or open questions?" The odds are that a fairly substantial number of people would not be able to define these two question types. However, if they are asked this question, they will be happy to choose. The point is that people will nct volunteer their ignorance of something, although they may admit it if you ask them. However, this caution goes beyond ignorance of an issue. Another problem is that the specialists wording the question may simply have an unusual command of their own language. Scientific jargon has been criticized. Perhaps overlooked is the fact that there are other kinds of jargon, too. The question askers have a responsibility to make themselves understood. One way of screening for individuals who do not have a basis for providing the information needed is to include one or two pure information questions. Plan to discard questionnaire returns from respondents who cannot answer the information questions correctly. However, our usual policy should be to throw out or revise items that are not understandable, rather than to throw out the responses of the people who can't understand the item.

Schaefer, Bavelas, and Bavelas (1980) developed a method to ensure that respondents would only be subjected to items that they could understand. The technique that they used is called "Echo." They developed items that were used in a performance rating scale. It would be possible to use the "Echo" technique in the development of survey items, too. Essentially, the "Echo" technique is a method for wording questionnaire items in the language of the respondents. A detailed procedure for using the "Echo" technique is available from J. B. Bavelas (1980).

The "Echo" Technique assumes that there are two separate populations in the development of questionnaire items. One population is the researchers, and the other population is the respondents. Phrasing of items needs to be in the language of the respondents. It requires content validation, i.e., confirmation that the content is understandable to the respondents. The "Echo" technique includes the development of a pool of items generated by a survey directed to the target population. The sample of potential respondents from the target population follows printed guidelines to write the items. Another sample from the target population is selected to sort items into categories. Part of this process includes concurrence by the members of the sample that the categories are mutually exclusive.

2. Ways of Measuring Item Difficulty

Various procedures exist for determining the difficulty or reading comprehension level of printed material. Such a discussion is, however, beyond the scope of this manual. Sources that may be consulted include:

a. Bavelas, J. B. (1980). In-house report for professionals and nonprofessional -- procedural details for the "Echo" technique. Victoria, British Columbia: University of Victoria, Department of Psychology.

b. Dale, E., & Chall, J. S. (1948). A formula for predicting readability. Educational Research Bulletin, 27, 11-20, 37-54.

c. Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32, 221-233.

d. Fry, E. (1968). A readability formula that saves time. Journal of Reading, 11, 513-516.

e. Lorge, I. (1944). Predicting readability. Teachers College Record, 45, 404-419.

f. Schaefer, B. A., Bavelas, J., & Bavelas, A. (1980). Using echo technique to construct student-generated faculty evaluation questionnaires. Teaching of Psychology, 7(2), 83-86.

g. Thorndike, E. L., & Lorge, R. (1944). The teacher's word book of 30,000 words. New York: Columbia University Press.

E. Length of Question/Stem

This section notes some considerations about the length of question stems. There is little research in this area to guide the questionnaire writer. See Section IX-C regarding questionnaire length.

1. It is sometimes desirable to break the question stem into two or more sentences when the sentence structure would otherwise be unnecessarily complex. For instance, one sentence can state the situation, and one can pose the question. Lengthy question stems that try to explain a complicated situation to the respondent should be avoided. If the respondents are not aware of the facts presented, they may become more confused or biased than enlightened, and their opinion would not mean much.

2. Longer open-ended questions do not necessarily produce a greater amount of and more accurate information than shorter ones. However, it may take more words to achieve a proper focus.

3. Questionnaire developers have a tendency to use long question stems with true-false questions when "True" is the correct answer. Respondents often detect and react to this tendency. Field test questionnaires, however, should make relatively little use of "True" and "False" response alternatives. These alternatives are more appropriately used when testing whether respondents have acquired a required proficiency level, for example, the ability to visually recognize a given type of enemy aircraft.

4. To obtain higher reporting levels by respondents when threatening questions are asked about their behavior, longer items may be best. Items with 30 or more words have achieved best results. Items with fewer words (less than 30) have not elicited reporting levels which were as high. One of the longer items had 49 words, and the content was about the use of drugs.

F. Order of Question Stems

There are two issues to consider regarding the order of question stems. The first has to do with the order of questions within a series of items that are designed to explore the same topic or subject matter or related subject matter areas. The second has to do with the order of different groups of questions when the groups deal with fairly separate topics or subject matter areas. For example, one group of questions may deal with factual items, while another may deal with attitudes. If items bearing on the same point are presented in succession, the respondent can proceed more readily through them. Thus, this is usually a desirable practice. An exception arises when one wishes to check the consistency of the respondents. To do this, two (or more) similar items are included, but at widely different points in the questionnaire.

1. Order of Questions Within a Series of Items

a. It is often recommended that the order of questions on an instrument be varied or assigned randomly to avoid one question contaminating another. The view is that the immediately preceding question or group of questions places the respondent in a "mental set" or frame of reference. For example, asking respondents a general question about their feelings regarding automobile exhaust pollution might influence responses to the question, "Do you prefer leaded or nonleaded gasoline?" Questionnaires are plagued by contextual effects attributed to item ordering. Respondents lacking in experience of the content area may change their responses as they progress through the questionnaire, since they may learn from previous items (order effects). This may damage the face validity of the responses to the initial items. Yet, the meaning of the items would be changed if they were separated from their topic areas. The current state-of-the-art for context effects suggests that all items which are interrelated by content area may be affected by context effects. There is currently no way to predict which items will have context effects.

b. Sometimes it is recommended that broad questions be asked before specific questions. The rationale for this approach is that the respondent can more easily and validly answer specific questions after having had a chance to consider the broader context. Also, asking the specific questions first could influence the response to the broader question. The quality of responses to questions on a questionnaire will be determined by the respondent's background and knowledge of the topic area. A series of specific questions (versus general questions) will provide information about whether the respondent understands the content of the questions. It should expose any logical inconsistencies in response patterns. Respondents with limited or no experience regarding the content area may deviate from the logical response pattern. Their answers to questions may change as they become more familiar with the topic through order effects. Early responses may not have face validity.

General and specific questions were empirically examined for order effects. The order of the questions did not appear to effect the way respondents marked the response alternatives. It is proposed that a stronger survey instrument may be provided by assigning general items first, followed by specific items on related topic areas. However, questions which are specific are preferred over general type questions. Contextual effects can be minimized by developing questions which are specific in content. Minimize the number of general questions.

c. The order of questions within a series of items will also depend upon whether filter questions are needed. A filter question is used to exclude respondents from a particular sequence of questions if those questions are irrelevant to them. For example, if a series of items were asked about different kinds of weapons, a "No" response to a question such as "Have you ever used the M14 rifle?" might be used to indicate that the respondent should skip the following question(s) about the M14.

When filter questions are used by an interviewer, they can reduce interviewing time. Clear branching instructions are imperative for the interviewer. Filter questions used to branch in mail surveys or group-administered questionnaires have the potential to increase non-response rate for questions which follow a branch. Items following a branch tend to receive a lower response rate. Response rate for individuals over 60 years of age are even lower. There are alternatives to branching such as the design of different questionnaires for different categories of respondents. The design of different questionnaires for different groups of respondents is illustrated in Figure VI-F-1.

Figure VI-F-1

Example of Bradley Fighting Vehicle Questionnaire for Multiple Groups

Questionnaires Designed for:

1. Driver
2. Track commander
3. Gunner
4. Other personnel

2. Order of Different Groups of Questions

a. There is usually a psychological or logical order in which to ask questions, so that the questionnaire flows smoothly from one topic to the next and the respondent is not shifted frequently from one topic to another and back again. However, when a shift is made from one topic to another, it should be apparent to the respondent.

b. It is usually recommended that more difficult or more sensitive questions be asked later in the questionnaire, possibly at the end.

c. One or more easy, non-threatening questions should probably be asked first to build rapport. They should be short and easy to understand and to answer. But they should not be irrelevant to the objectives of the questionnaire. Verbal efforts to build rapport by the questionnaire administrator seem preferable to using questionnaire content to accomplish this task.

G. Number of Response Alternatives

The following sections consider number of response alternatives to use in multiple choice, rating scale, and forced choice items: Section VI-C-3 - formulation of response alternatives; Section VI-H - order of response alternatives; Chapter VII - response anchoring; Chapter VIII - order of perceived favorableness of words and phrases.

One of the basic issues in the use of rating questions or attitude scales is the determination of the optimum number of responses, alternatives or categories. In questionnaire construction, researchers have investigated the utility of having a scale with a greater or smaller number of scale points. Over the years, there have been diverse recommendations on the proper number of scale points or categories to use in questionnaire construction. Investigations have indicated that reliability was optimum for scale points of 2, 5, 10, 11, 20, and 25. Some recent research has proposed the use of a range of scale points between 2 and 10. The reason for concern with the number of response alternatives is due to the belief that a "coarse" scale with too few response alternatives may result in a loss of information concerning the respondents' discrimination powers. It may reduce the respondents' cooperation in rating, as a coarse scale "forces" judgments and thereby irritates some respondents. An extremely "fine" scale, with too many response alternatives, may go beyond the respondents' powers of discrimination, be excessively time consuming, or difficult to score.

1. Number of Response Alternatives with Multiple Choice Items

No firm rules can be established regarding the number of response alternatives to use with multiple choice items. It depends in a large part upon the question being asked, and the number of answers logically possible. The following considerations, however, may be noted:

a. There is some evidence that dichotomous items (items with only two response alternatives) are statistically inferior to items with more than two response alternatives.

b. Dichotomous items are easier to score than nondichotomous items, but they may not be accepted as well by the respondent.

c. A good nondichotomous multiple choice item usually cannot be written as a set of separate dichotomous items.

d. Consideration should be given to the prospect that many response alternatives may make a questionnaire unduly time consuming.

e. The number of choices logically possible or desirable should constitute an upper limit on the number of response alternatives used for an item.

f. Non-existent response alternatives may be checked by the respondent if an answer sheet is used which has more spaces than there are alternative answers; e.g., the answer sheet has five spaces for each question, but some questions have fewer than five alternatives.

2. Number of Response Alternatives with Rating Scale Items

Authorities in psychometrics contend that the optimal number of response alternatives to employ with rating scales is a matter for empirical determination in any situation. They also suggest that considerable variation in number around the optimal number changes reliability very little. These conclusions seem to be supported by the available research literature. Although rules regarding the number of response alternatives to use with rating scales cannot, therefore, be firmly established, the following issues can be considered.

a. The effects of increasing or decreasing the number of response alternatives for a question cannot be generally specified with certainty. Increasing the number of response alternatives does not necessarily increase reliability, and there is no consistent relationship between the number of response alternatives and validity.

b. J. P. Guilford (in Psychometric methods. New York: McGraw-Hill, 1954) reported that seven response alternatives is usually lower than optimal, and it may pay in some favorable situations to use up to 25 scale divisions. Others believe that seven steps or five is optimal. Some believe that five should be used for single or unipolar (one direction)) scales, nine for double or bipolar scales. Many practitioners consistently use five-point scales. Sometimes a nine-point hedonic (pleasure) scale is recommended for food items, and a six-point scale for other uses.

c. The number of response alternatives to use is often determined on the basis of the degree of discrimination required. For example, a nine-point scale may sometimes (but not always) give greater discrimination than a three-point scale. Increases in reliability tend to level off after seven scale points, and there is no apparent advantage in using a large number of scale points.

d. Psychologists with considerable experiences in military operational field testing feel that anything more than five alternatives is too great a number for many junior enlisted personnel to discriminate among. More nonresponses are obtained, and the discrimination power of answered items is not increased.

e. Questionnaire administration time is probably a function of the number of response alternatives.

f. There is some evidence that increasing the number of response alternatives seems to decrease the number of nonresponses and uncertain responses (e.g., "Cannot decide").

g. In addition to the response alternatives representing the rating scale continuum, it may be necessary to add alternatives such as "No opinion" or "Did not experience."

h. Scoring and data analysis considerations may affect the selection of the number of response alternatives. If Chi square tests are sufficient, two or three response alternatives might be adequate. However, if two or three response alternatives are used when nonparametric rank order correlations are employed, substantial "ties" on ranks will result. If parametric statistics are to be employed, more alternatives are usually better, because of the assumption of continuous distributions or interval scale properties.

i. In some situations, fully-labeled scales may discriminate better than only end-anchored scales. Responses to fully-labeled scales may be less skewed than responses to only end-anchored scales.

3. Number of Response Alternatives with Forced Choice Items

A number of different forced choice item formats have been used, such as the following:

a. Two phrases or statements per item, both favorable or both unfavorable, choose the more descriptive or the least descriptive.

b. Three statements per item, all favorable or unfavorable, choose the most and least descriptive statements in each item.

c. Four statements per item, all favorable, choose the two most descriptive statements.

d. Four statements per item, all favorable, choose the most and least descriptive statements.

e. Four statements per item, two favorable and two unfavorable, choose the most and least descriptive statements.

f. Five statements per item, two of which were favorable, one neutral, and two unfavorable in appearance, choose the most and least descriptive.

The evidence is not clear, but three or four statements per item may be preferable to two. One study concluded that the format described in "c" above was superior to the others. It was most bias resistant, yielded consistently high validities under various conditions, had adequate reliability, and was one of the best received by respondents.

H. Order of Response Alternatives

1. General Considerations

The experimental evidence on the effect that the order of presentation of response alternatives for a question has on a subject's choice of response is inconclusive and contradictory. Varying conclusions include:

a. Respondents have a tendency to select the first response alternative in a set more than the others.

b. With multiple choice questions, there is a tendency to choose answers from the middle of the list, if the list consists of numbers. Answers were selected from either the ton or bottom of the list, if the alternatives were fairly lengthy expressions of ideas.

c. Longer items produced responses that were closer to the center of the response scale. Shorter items yielded more positive responses.

d. Poorly motivated respondents tend to select the center or neutral alternatives with rating scale items.

e. Fully-labeled response alternatives yielded less skewed response distributions than only labeling the endpoints.

f. On items about which respondents feel strongly, the order of alternatives makes no difference. On items about which the respondent does not feel strongly, most will tend to check the first alternative.

g. Items that were positively worded received higher mean responses than negatively-worded responses.

h. The positive pole of rating scale response alternatives should be presented first since this will improve the reliability of the responses. However, it is important to realize that reliability may increase while validity decreases.

i. Placement of either the positive or negative endpoint at the left-hand side of the semantic differential scale was not associated with response style.

j. Semantic differential scales were found to confound trait self-descriptions with socially desirable responses on clinical instruments. When a socially undesirable adjective anchor was presented first, subjects had a tendency to select adjectives which were opposite in desirability.

Test item form biases are discussed in Section XII-B.

2. Suggested Order for Multiple Choice Items

The following suggestions are offered regarding the order of multiple choice items:

a. When the response alternatives have an immediate apparent logical order (e.g., they all relate to time), they should be put in that order.

b. When the response alternatives are numerical values, they should in general be put in either ascending or decreasing order.

c. When the response alternatives have no immediately apparent logical order, they should generally be put in random order.

d. Alternatives such as "none of the above" or "All of the above" should always be in the last position.

e. Alternate questionnaire forms (e.g., where the order of alternatives is reversed on half of the forms) are often desirable.

f. More abstract types of questions minimize order effects by developing questions which are specific in content instead of general,

3. Suggested Order of Rating Scale Items

Since rating scales call for the assignment of objects along an assumed continuum or in ordered categories along the continuum, it follows that the response alternatives must be in order from "high" to "low" or "low" to "high," with the choice of words for "high" and "low" (the endpoint labels) depending upon the continuum being used. For example, for the continuum satisfactory-unsatisfactory, item 1 in Figure VI-H-1 uses the "high" to "low" order, while item 2 uses the order "low" to "high."

Figure VI-H-1

Example of Rating Scale Item
with Alternate Ordering of Response Alternatives

1. The M16 rifle is:

_____ very satisfactory.

_____ satisfactory.

_____ borderline.

_____ unsatisfactory.

_____ very unsatisfactory.

2. The M16 rifle is:

_____ very unsatisfactory.

_____ unsatisfactory.

_____ borderline.

_____ satisfactory.

_____ very satisfactory.

Many practitioners use the "high" to "low" order. If one has reason to believe that the order of the response alternatives makes a difference, or wishes to make certain that they do not, then the use of alternate questionnaire forms is recommended. Each alternate form should list the response alternatives in a different order. The "good" or "high" end of the scales should be at the same end of each scale for all items in a given questionnaire form, but the order should normally be reversed on 50% of the forms. For example, the order shown in item 1 in Figure VI-H-1 would be used on half of the forms; the order shown in item 2 on the other half. (Normally, there would be only two questionnaire forms, one with each order, but at times alternate forms are also needed for other purposes. Hence, there may be more than two.)

Chapter VII: Response Anchoring

A. Overview

This chapter addresses the "anchoring" of rating scale responses; that is, the words used to define some or all of the response alternatives. Section VII-B shows various types of response anchors, while Section VII-C discusses anchored versus unanchored scales. The amount of verbal anchoring is the topic of Section VII-D, while some procedures for the selection of verbal scale anchors are presented in Section VII-E. Finally, Section VII-F discusses balanced versus unbalanced scales.

It should be noted that Section VI-C 3 discussed the formulation of response alternatives, while the number and order of response alternatives are the topics of Sections VI-G and VI-H, respectively. The order of perceived favorableness of words and phrases is discussed in Chapter VIII.

B. Types of Response Anchors

There are a number of different types of response anchors that can be used with rating scale items. Some have been shown as examples in other chapters, such as Section VI-D. Nine types of response anchors are shown in Figure VII-B-1. The first shows the original form of the semantic differential. It is a combination graphic and verbal scale. Respondents were instructed to place an "X" at a place on the line that would represent their attitude. The use of verbal anchors with a -5 through +5 numerical continuum is shown in item 2 of Figure VII-B-1. Item 3 shows verbal anchors used with a 1 through 11 numerical continuum. There is evidence that variables studied by behavioral scientists are continuously distributed, even though the measuring instruments yield discrete scores. These scores are approximations of the supposedly continuous variables. A combination verbal and numerical continuum (series) is shown in item 4, while a verbal and alphabetical continuum is shown in item 5. Item 6 is similar to item 5 since it too is a verbal continuum. This item lacks the alphabetical and numerical response anchors associated with other verbal anchors. Item 7 is a typical Likert rating scale that calls for a verbal rating to a directional statement that may be phrased either positively or negatively. An example might be "The Modern Volunteer Army places too much emphasis on extrinsic factors (such as beer in the barracks) as opposed to intrinsic, job related factors (such as pay or supervision)." Item 8 is constructed on a continuous scale to obtain more discrimination along the scale line, and it is verbally anchored. Item 9 is one behavioral anchor from a set of nine scale points. This particular behavioral anchor has a scale point of four.

Conflicting empirical evidence exists regarding the reliability of scales with verbal anchors and verbal response alternatives so that neither is superior to that of purely numerical scales. Some feel that adding verbal anchors to a scale will increase reliability. Recent research in ergonomics and other related applications indicates that either numerical response alternatives or verbal response alternatives are psychometrically acceptable. If verbal anchors are used, be sure they are properly developed.

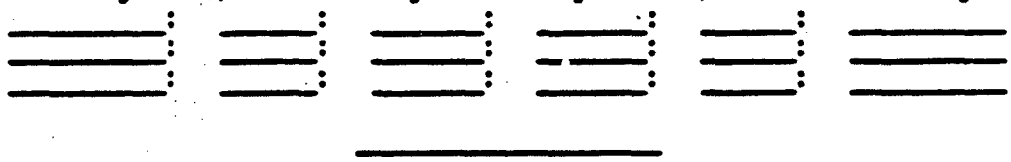
Figure VII-B-1

Types of Response Anchors

1. Combination graphic and verbal scale.

Strong _____: _____: _____: _____: _____: _____: _____: _____: _____ Weak

Extremely Quite Slight Slight Quite Extremely



LOW HIGH

2. Verbal anchors with a -5 through +5 numerical continuum (series).

Definitely dislike ... Definitely like

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

3. Verbal anchors with a 1 through 11 numerical continuum (series).

Definitely dislike ... Definitely like

1 2 3 4 5 6 7 8 9 10 11

4. A verbal and numerical continuum (series).

Dislike completely	Dislike somewhat	Dislike a little	Neither like nor dislike	Like a little	Like somewhat	Like completely
1	2	3	4	5	6	7

5. A verbal and alphabetical continuum (series).

Below Average	Average	Above Average	Well Above Average	Outstanding
(A)	(B)	(C)	(D)	(E)

Figure VII-B-1 (Cont.)

Types of Response Anchors

6. A verbal continuum (series).

Below average	About average	A little better	A lot better	One of the best	None better

7. A verbal continuum (series). (Likert rating scale)

__Agree strongly __Agree __Undecided __Disagree __Disagree strongly

8. Combination verbal and continuous (series) scale.

Attribute

negative neutral positive

9. Combination behavioral anchor and numerical scale point.

Scale Point	
4	Many troops in this unit would leave the post as quickly as possible after duty hours to avoid doing any extra work.

C. Anchored Versus Unanchored Scales

A number of studies have been conducted on the topic known as "anchoring effects." Unfortunately, the research evidence is contradictory as to whether anchored or unanchored scales should be used. It has been noted that unanchored scales may well be anchored by the question stem, so that the response alternatives may not have to be. When only one end of a scale is anchored, some studies have found a tendency for respondents to move toward that extreme. But other studies have found the opposite tendency. At least one study found that judgment and response time is decreased with anchoring. In practice, then, it is usually best to use anchored scales.

D. Amount of Verbal Anchoring

Obviously, the amount of verbal anchoring of a rating scale item can vary. It can be anchored at the center, or on the ends, or both, or at many points on the entire continuum. There is some evidence that more descriptive data can be obtained with more anchoring, and that greater scale reliability is achieved with added verbal anchoring. In one study, scales labeled at only the extreme endpoints resulted in responses that were skewed toward the positive end of the scale. Scales with verbal descriptors for all response alternatives may also be better predictors of behavior. On the other hand, adding examples to definitions does not seem to help much. (See also Section VI-G regarding the number of response alternatives to employ.) Fully labeled scale points may encourage a more balanced response distribution.

E. Procedures for the Selection of Verbal Scale Anchors

Some guidance can be offered regarding the selection of verbal scale anchors. See also Chapter VIII.

1. Pretests for the selection of verbal anchors are valuable in building scale content. Rather than employing anchors which seem appropriate, anchors should preferably be selected by respondents similar to those who will be participating in the study.

2. Scale endpoints that are unrealistically extreme, such that few if any respondents would select them, should be avoided. For example, it may be seldom that "Never" or "Always" apply. The use of "Rarely" and "Usually" may be more appropriate. There are instances, however, where extreme statements are realistic. The decision here often requires experience with what is being rated.

3. Analysis of data is normally facilitated if verbal scale anchors selected for rating scales are of equal distance from each other in terms of scale values. See, however, Chapter VIII.

4. Scales can be anchored by examples of expected behavior based upon observations of behavior. There are a wide variety of behavioral scales using variations of the Smith and Kendall format. These scales use behavioral anchors constructed from critical incidents. Procedures for establishing behavioral anchors may be found in the following references.

   a. Bernardin, H. J., La Shells, M. B., Smith, P. C., & Alvares, K. M. (1976, February). Behavioral expectation scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 61(1), 75-79.

   b. Borman, W. C., & Dunnette, M. (1975). Behavior-based versus task-oriented performance ratings: An empirical study. Journal of Applied Psychology, 60, 561-565.

   c. Finley, D. M., Osborn, H. G., Dubin, J. A., & Jeanneret, P. R. (1977). Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 30, 659-669.

   d. Fivars, G. (1975). The critical incident technique: A bibliography. JSAS Catalog of Selected Documents in Psychology, 5, 210.

   e. Landy, F. J., & Barnes, J. L. (1979). Scaling behavioral anchors. Applied Psychological Measurement, 3(2), 193-200.

   f. Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

F. Scale Balance, Midpoints, and Polarity

1. Balanced Versus Unbalanced Scales

Historically, balanced scales have been preferred by researchers. A scale is balanced when it has a number of positive response alternatives equal to the number of negative alternatives, regardless of the presence or absence of an "indifferent," neutral, or mid-scale category. A "Don't know" response alternative, if present, is not considered to be part of the scale, so is not counted when deciding if the scale is balanced. See the examples of balanced and unbalanced scales in Figure VII-F-1. Unbalanced scales may be employed if pretest results indicate that many respondents will be choosing extreme response alternatives at one end of a scale, producing a skewed distribution of responses rather than the statistically expected normal distribution around the mean attitude. To reduce the piling up of responses at one end of a scale, - or, to add to your ability to discriminate among responses in that region - the scale is made unbalanced by adding more response alternatives on the side of the scale where the piling is likely to occur. This practice tends to spread the distribution of responses more evenly along the scale continuum.

In cases where one has no advance information or other basis for expecting responses to be largely one-sided, it is normally desirable to have an equal number of positive and negative response alternatives; i.e., a balanced scale.

2. Midpoints

Scales may or may not include a midpoint or mid-scale response alternative. This does not affect their classification, but does affect their response distributions. There is research evidence that when a middle position is offered on a scale, there is a shift in the distribution of ratings. Up to 10-20% or more of the ratings may shift into the midpoint causing a decline in the polar positions. Even so, questionnaires that have response distributions that include a midpoint yield similar distributions to those without the midpoint. The inclusion or exclusion of a midpoint probably won't influence the response distribution that much as long as there are at least five scale points.

As examples, items 1c, 2a, and 3 in Figure VII-F-1 show scales with no mid-scale point. One might exclude the mid-scale point for items where it is judged that respondents ought to have a sufficient basis for being pro or con, and where one desires to force respondents away from an "on the fence" position. Bipolar scales should be balanced in terms of the degree of extremeness denoted by the endpoint anchors. For example, if "Never" is used, then "Always" should be used as the opposite endpoint.

Figure VII-F-1

Examples of Scale Balance, Midpoints, and Polarity

1. Balanced bipolar scales.

a. Very progressive
Progressive
Moderately progressive
Neither progressive nor conservative
Conservative
Very conservative

b. Effective
Fairly effective
Borderline
Fairly ineffective
Ineffective

c. Very effective
Somewhat effective
Somewhat ineffective
Very ineffective

d. Very satisfied
Satisfied
Borderline
Dissatisfied
Very dissatisfied

2. Unbalanced bipolar scales.

a. Enthusiastic
Extremely favorable
Very favorable
Favorable
Fair
Poor

b. Quite good
Rather good
Somewhat poor
Rather poor
Quite poor
Very poor

3. Unbalanced Scale (unipolar).

Very much
Much
Some
A little
None

3. Polarity

Scales may be bipolar or unipolar. Item 3 in Figure VII-F-1 illustrates a unipolar scale. Its basic feature is that it represents the thing being assessed as having from none to a maximum - with n steps in between - of some property. The question of balance only arises for bipolar scales. Many a bipolar scale could be redesigned as a unipolar scale. Instead of item 1c in Figure VII-F-1, one's question about effectiveness (not given) could have been followed by this unipolar scale of effectiveness: maximum effectiveness, great effectiveness, moderate effectiveness, slight effectiveness, and no effectiveness.

Semantic preferences may determine whether the questionnaire writer uses bipolar or unipolar scales.

Chapter VIII: Empirical Bases for Selecting Modifiers for Response Alternatives

A. Overview

When constructing a questionnaire, it is often necessary to select adjectives, adverbs, or adjective phrases to use as response alternatives. The words selected for response alternatives should be clearly understood by the respondents to the questionnaire. and they should have precise meaning. There should be no confusion among respondents as to whether one term denotes a higher degree of favorableness or unfavorableness than another.

There is no need to guess which phrases or words are the best to use as response alternatives. Many studies have been conducted in order to determine the perceived favorableness of commonly used words and phrases. These studies have determined scale values and variances for words and phrases which can be used to order the response alternatives. In some of the studies, ambiguous words and words that are not appropriate to use as response alternatives have been identified.

The results of these studies and the experience of questionnaire designers have been incorporated into this chapter in order to offer guidelines and suggestions to be used in selecting response alternatives. This chapter includes lists of words and procedures to use in selecting response alternatives. Many lists of phrases with mean scale values and standard deviations are presented. The scale values are given for the purpose of selecting response alternatives, not for the purpose of assigning scale values to response alternatives for data analysis purposes.

Section VIII-B discusses things to consider in selecting response alternatives; Section VIII-C covers the selection of response alternatives denoting degrees of frequency; Section VIII-D, the selection of response alternatives using order of merit lists of descriptor terms; Section VIII-E, the selection of response alternatives using scale values and standard deviations. Section VIII-F includes sample sets of response alternatives.

Scale values, standard deviations, and interquartile ranges reported in this chapter have been taken from data presented in the following studies:

1. Altemeyer, R. A. (1970). Adverbs and intervals: A study of Likert scales. Proceedings of the Annual Convention of the American Psychological Association, 5(pt. 1), 397-398.

2. Backstrom, C. H., & Hurchur-Cesar, G. (1981). Survey research. New York, NY: John Wiley & Sons.

3. Beltramini, R. F. (1982). Rating-scale variations and discriminability. Psychological Reports, 50, 299-302.

4. Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and the number of categories on the scale. Journal of Applied Psychology, 37, 38-41.

5. Boote, A. S. (1981). Reliability testing of psychographic scales. Journal of Advertising Research, 21(5), 53-60.

6. Cliff, N. (1959). Adverbs as multipliers. Psychological Review, 66, 27-44.

7. Dodd, S. C., & Gerberick, T. R. (1960). Word scales for degrees of opinion. Language and Speech, 3, 18-31.

8. Dolch, N. A. (1980). Attitude measurement by semantic differential on a bipolar scale. The Journal of Psychology, 105, 151-154.

9. Gividen, G. M. (1973, February). Order of merit- descriptive phrases for questionnaires. Unpublished report, available from the ARI Field Unit at Fort Hood, TX.

10. Innes, J. M. (1977). Extremity and "don't know" sets in questionnaire response. British Journal of Social Clinical Psychology, 16, 9-12.

11. Ivancevich, J. M. (1980). Behavioral expectation scales versus nonanchored and trait rating systems: A sales personnel application. Applied Psychological Measurement, 4(1), 131-133.

12. Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics: An experimental investigation. Journal of Applied Psychology, 39, 31-36.

13. Mathews, J. L., Wright, C. E., & Yudowitch, K. (1975, March). Analysis of the results of the administration of three sets of descriptive phrases. Palo Alto, CA: Operations Research Associates.

14. Mathews, J. L., Wright, C. E., Yudowitch, K. L., Geddie, J. C., & Palmer, R. L. (1978, August). The perceived favorableness of selected scale anchors and response alternatives (Technical Paper 319). Palo Alto, CA: Operations Research Associates, and Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (DTIC No. AD A061755)

15. Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image: An evaluation. Journal of Marketing Research, 16(1), 80-87.

16. Mosier, C. I. (1941). A psychometric study of meaning. Journal of Social Psychology, 13, 123-140.

17. Myers, J. H., & Warner, W. G. (1968). Semantic properties of selected evaluation adjectives. Journal of Marketing Research, 5, 409-412.

18. Presser, S., & Schuman, H. (1980). The measurement of a middle position in attitude surveys. Public Opinion Quarterly, 44(1), 70-85.

19. Reynolds, T. J., & Jolly, J. P. (1980). Measuring personal values: An evaluation of alternative methods. Journal of Marketing Research, 17, 531-536.

20. Schuman, H., & Presser, S. (1981). Questions and answers in attitude surveys: Experiments on question form, wording, and context. New York: Academic Press, Inc.

21. U.S. Army Test and Evaluation Command (1973). Development of a guide and checklist for human factors evaluation of Army equipment and systems. U.S. Army Test and Evaluation Command (TECOM).

B. General Considerations in the Selection of Response Alternatives

There are several ways of selecting response alternatives. These ways are dependent on the purpose of the questionnaires and/or on the way the data will be analyzed. There are specific considerations when selecting response alternatives for balanced scales, when selecting response alternatives with extreme values, and when developing equal interval scales. There are also general things to consider in the selection of any response alternative.

In some cases, it is desirable to select response alternatives on more than one basis. For example, mutually exclusive phrases may be selected also on the basis of parallel wording.

1. Matching the Question Stem

Descriptors should be selected to follow the question stem. For example, if the stem asks for degrees of usefulness, descriptors such as "Very useful" and "Of significant use" should be used. In some cases, this may mean rewording the question stem so that appropriate response alternatives can be selected.

2. Mixing Descriptors

Descriptors on different continuums should usually not be mixed. For example, "Average" should never be used with quantitative terms or qualitative terms such as "Excellent" or "Good" (since "average" performance for a group may very well be excellent or good or even poor). If the descriptors are selected for use with a question stem asking about satisfactory or unsatisfactory, the word "Satisfactory" or "Unsatisfactory" (or a synonym) should normally be in every response alternative, except perhaps for a neutral response alternative.

Some experts go as far as to say that the wording of the response alternatives should be parallel for balanced scales. For example, if the phrase "Strongly agree" is used, then the phrase "Strongly disagree" should also be used. By reviewing some of the studies that have determined scale values for descriptors, it can be seen that some pairs of parallel phrases are not equally distant from a neutral point or from other phrases in terms of their scale values. Hence, parallel wording may not always provide equally distant pro and con response alternatives, although they may be perceived as symmetrical opposites.

Using descriptors from one continuum or descriptors with parallel wording for a given questionnaire item has advantages. The advantages are that the response alternatives will usually fit the stem better, and they will be parallel to each other in meaning and appearance.

3. Selecting Response Alternatives with Clear Meaning

Some words are difficult for respondents to use in answering questions. This difficulty may be the result of the respondent being ignorant of the meaning of the word, or not being able to rate the word in terms of degrees on specific scales. Such words should not be used as response alternatives. Some studies asked the respondents to indicate which words they were unable to rate. Table VIII-B-1 lists examples of words that were unrateable by subjects.

Table VIII-B-1

Words Considered Unrateable by Subjects

Phrase	Phrase
Adverse	Noxious
Appalling	Peerless
Base	Satiating
Despicable	Seemly
Expedient	Superlative
Fit	

From: Mosier 1941a.

Some words appear to have two or more distinct meanings. When these words are rated on a continuum of favorableness-unfavorableness, many respondents will mark one part of the scale, while the other respondents will mark a different place on the scale. This depends on how they interpret the words, and may result in a bimodal distribution. Such words also should not be used as response alternatives. A list of words evoking bimodality of response is given in Table VIII-B-2.

Table VIII-B-2

Words Evoking Bimodality of Response

Word(s)	Word(s)
Acceptable	Irresistable
Amazing	Normal
Bearable	Tempting
Completely indifferent	Unfit
Extremely indifferent	Unspeakable
Highly indifferent	Unusually indifferent
Important	Very indifferent
Indifferent	Very, very indifferent
Indispensable	

From: Mosier 1941a

4. Selecting Nonambiguous Terms/Descriptors

Some descriptors are more ambiguous than others. The more ambiguous the descriptor, the more varied the respondents' interpretations of the degree of favorableness denoted by the descriptor. The ambiguousness of a descriptor is measured by the variability of responses given to the item. One measure of variability is the standard deviation. When available, standard deviations (SD) are given with scale values in this chapter. Another measure used to show variability is the interquartile range. This measure is indicated in this chapter with scale values only when the standard deviations were unavailable.

It is most desirable to select terms with small ranges or small standard deviations, as they will have less ambiguous meaning to respondents. Also, selecting a term with a small standard deviation decreases the chances of the meaning of the term overlapping with the meaning of neighboring terms.

5. Selecting Response Alternatives

When balanced scales with two, three, four, or five descriptors are sufficient for describing the distribution of respondents' attitudes or evaluations, the questionnaire writer can compose them quite satisfactorily by using a term and its literal opposite (effective vs. ineffective; pleasing vs. unpleasing) for two of the terms. A more extreme pair can be produced by using "Very" to modify these two terms.

The first of several intended studies of how people rate/order terms that might be used for rating scale descriptors was conducted by Operations Research Associates and ARI just prior to the writing of this manual. Its results may assist questionnaire developers who need unbalanced scales or scales with more than five descriptors. In the study, each of 100 Army personnel was asked to assign a scale value ranging from -5 (most negative) to +5 (most positive) to each term in three different sets of terms, totaling over 100 descriptors.

Tables VIII-B-3 and VIII-B-4 give samples of descriptors from this study for which mean scale values and standard deviations have been calculated. The list in Table VIII-B-3 was derived by first selecting the descriptor with the largest positive mean. The next descriptor selected has a mean that is at least one standard deviation lower. The implication of the gap of one standard deviation is that not more than 16% of the people would have assigned a lower scale value to the first descriptor than they did to the second descriptor, and vice versa. To this extent, the raters disagreed on the ordering of these two terms when rating about 50. The third descriptor on the list has a mean scale value yet another standard deviation lower. This process was repeated until the descriptor with the lowest mean scale value was selected. A descriptor was not used if its standard deviation was greater than 1.000.

The list on Table VIII-B-4 was constructed again by skipping at least one standard deviation between adjacent terms; however, the starting point was at the middle, with the word "Neutral."

Use of Table VIII-B-3 as a 10-descriptor unbalanced scale is not highly recommended. If one wanted a nine-descriptor scale, one could use the four adverbs appearing in front of "Acceptable" in the table in that same location, and also use them in front of "Unacceptable" in reverse order to create a semantically balanced and ordered scale. Or, one could use the five adverbs, now shown below "Neutral," both above and below "Neutral" to create an 11-descriptor scale of acceptability (or effectiveness, or satisfactoriness, etc.). "Neutral," however, may not be a suitable midpoint term here as the respondent who has neutral feelings (i.e., does not know or does not care) might check this response, whereas the term "Neutral" is intended to specify, for example, a midpoint between "barely acceptable" and "barely unacceptable."

Table VIII-B-3

Sample List of Phrases
Denoting Degrees of Acceptability

Phrase	Mean	SD
Wholly acceptable	4.725	.563
Highly acceptable	4.040	.631
Reasonably acceptable	2.294	.722
Barely acceptable	1.078	.518
Neutral	.000	.000
Barely unacceptable	-1.100	.300
Rather unacceptable	-2.020	.836
Substantially unacceptable	-3.235	.899
Highly unacceptable	-4.220	.576
Completely unacceptable	-4.900	.361

From: Mathews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-B-4

A Second Sample List of Phrases
Denoting Degrees of Acceptability

Phrase	Mean	SD
Very, very acceptable	4.157	.825
Largely acceptable	3.137	.991
Mildly acceptable	1.586	.700
Sort of acceptable	.940	.645
Neutral	.000	.000
Barely unacceptable	-1.100	.300
Rather unacceptable	-2.020	.836
Substantially unacceptable	-3.235	.899
Highly unacceptable	-4.294	.535
Completely unacceptable	-4.900	.361

From: Mathews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

While the scale values from the studies cited are useful, further refinement is possible. That is, once having selected a candidate scale (set of descriptors), one could then conduct another study to determine if relevant judges would assign scale values indicating equal intervals (among means) for the terms on the candidate scale.

6. Selecting Descriptors for Endpoints

Once the decision has been made as to how extreme the endpoints of a scale should be (see Section VII-E 4), the descriptors should be selected accordingly. If extreme endpoints are desired, descriptors that have extreme meanings should be selected. One guideline that can be used in selecting these descriptors is to use those that have the highest and lowest scale values. Another guideline is to review the descriptors in terms of their apparent meanings. If less extreme endpoints are desired, descriptors that do not have extreme scale values and that do not have the apparent extreme meanings should be selected.

There has been conflict about whether fully-labeled scales are psychometrically superior to scales labeled only at the endpoints. Some evidence supports fully-labeled scale points which appear to produce response distributions that are less skewed.

7. Selecting Midpoint Responses

Whether a middle response alternative is included or excluded on a scale won't make that much difference as long as the number of scale points is at least five and not more than eleven. For respondents who have a weak opinion, eliminating the middle alternative will force a response toward either end of a bipolar scale. Including the middle alternative may increase differentiation of response, and may be useful for individuals who have a strong opinion on the topic. Overall, response distributions for scales that include the middle alternative look about the same as the distributions without the middle response alternative. The decline in responses to endpoints of the scale accounts for the shift in response when a middle alternative is offered.

Identification and selection of the label for the middle alternative is dependent on the overall selection of the other response alternatives. Different populations will perceive response alternatives with divergent perceptions. The means and variances for agreement on the semantic meaning of response alternatives will vary by population. To identify appropriate response alternatives, a sample from the target population could rate response alternatives for agreement for semantic meaning. Rating response alternatives by themselves may produce different results than rating response alternatives in conjunction with the item stem.

In selecting a descriptor for a midpoint response, it is necessary to use a descriptor that is neutral (neither positive nor negative) in meaning. Some of the commonly used midpoints do not appear as neutral as might be expected to some respondents.

Table VIII-B-5 lists several candidate midpoint terms with their scale values and standard deviations. This list may be helpful in selecting midpoint responses.

Table VIII-B-5

Candidate Midpoint Terms' Scale Values and Standard Deviations as Determined by Several Different Studies

Term	Mean Scale Value	SD	Theoretical Middle Scale Value
About average	3.77	.85	3.50
Acceptable	.73	.66	.00
Acceptable	11.12	2.59	10.00
Acceptable	2.39	1.46	.00
All right	10.76	1.42	10.00
Average	3.08	--	3.00
Average	.86	1.08	.00
Average	10.84	1.55	10.00
Borderline	-.02	.32	.00
Borderline	.00	.20	.00
Borderline	-.06	.31[a]	.00
Doesn't make any difference	2.83	3.73[a]	5.00
Don't know	4.82	.82[a]	5.00
Fair	6.50	--	5.50
Fair	.78	.85	.00
Fair	9.52	2.06[a]	10.00
Fair	4.96	.77[a]	5.00
Neutral	.00	.00	.00
Neutral	.02	.18	.00
Neutral	9.80	1.50	10.00
Neutral	10.18	2.01	10.00
Normal	6.70	1.43	6.00
Ordinary	6.50	1.43	6.00
O.K.	.87	1.24	.00
O.K.	10.28	1.67	10.00
So-so	10.08	1.87	10.00
Undecided	4.76	3.73[a]	5.00

[a]Interquartile range shown rather than the standard deviation

Words commonly used for midpoint responses are discussed below:

a. Average.

"Average" should never be used in conjunction with adjectives such as "Excellent," "Good," etc. "Average" has no meaning when used with these words. For example, "Average" performance may be superior or it may be completely unsatisfactory. Furthermore, most evaluators do not have the experience or competence to even know what an "Average" performance is. Typically, when "Average" is used on a field test evaluation form, only 5% or 10% of responders rate the subject as below average and 30% or 40% rate it above average. The data from such a question indicate that the response alternatives are not well formulated. Therefore, as a general rule, it is usually inappropriatee to use any term of "Average" in a questionnaire, and it is always inappropriate to use "Average" in conjunction with phrases such as "Excellent," "Good," "Poor," etc.

If "Average" is used, it should be with extreme care and only when one is interested in comparing performances or items with each other. It should not be used when one desires to find out how "good" or how "bad" an item or performance is. Significantly above average performance may be extremely unsatisfactory.

b. No opinion.

"No opinion" is unacceptable as a mid-scale term, as it usually denotes that a person has no opinion due to lack of knowledge or due to not having thought about an issue. "No opinion" can be used as a response alternative if it represents a specific type of information that is wanted.

c. Neutral.

"Neutral" is considered as a less desirable mid-scale term to use than "Borderline." Although every respondent in the study gave the term zero, the meaning on a questionnaire is not clear (see page VIII-B 4). Two out of 52 respondents indicated it was unrateable. In another study, "Neutral" had a mean scale value of .02 and a standard deviation of .18. Because of the ambiguity of meaning of "Neutral" (e.g., feeling of the respondent versus midpoint alternative), it is not recommended that it be used as midpoint on most questionnaires.

d. Marginal.

"Marginal" is sometimes used as a midpoint response alternative. Interviews with test subjects indicated that the term "Marginal" in most cases had a meaning of above "Borderline" or still satisfactory, but very close to being unsatisfactory. Hence, indications are that there may be more desirable terms to use than "Marginal."

e. Borderline.

"Borderline" is preferred by some experts as a midpoint response. In an administration to Fort Hood soldiers of over 1,500 questionnaires using the term "Borderline" as a midpoint, there was not one instance of reported confusion among those completing the questionnaires. However, there were times when "Borderline" had a larger standard deviation than "Neutral." (Again, "Neutral" by definition implies zero to most persons, but its frame of reference is ambiguous).

f. Uncertain.

"Uncertain" is unacceptable as a midpoint term, as it implies that with additional knowledge or thought a decision could be made that would fall into one of the other categories.

g. Undecided.

"Undecided" is also unacceptable as a mid-scale term for the same reasons as "Uncertain."

h. Neither agree nor disagree.

"Neither agree nor disagree" and similar descriptors written in this form may be used as midpoint responses. They have the advantage of paralleling the rest of the descriptors in the set, and they denote a position exactly in the middle of the endpoints. This term, like "Neutral," can also imply uncertainty, indecision, or a lack of knowledge rather than a firm knowledge that it represents a midpoint.

i. No effect.

"No effect" may be employed as a midpoint term when it is used with a set of descriptors to measure the type of effect that an activity will have. For instance, it can be used on a continuum from beneficial to detrimental.

j. Ordinary.

"Ordinary" should not be used as a mid-scale term. In one study, use of the term "Ordinary" as the mid-scale value resulted in a marked skewing of responses at the low end of the scale. This resulted from the common use of "Ordinary" to imply inferiority.

k. Fair.

"Fair" should not be used as a mid-scale term. In one study, the median scale value for "Fair" was a full point above the mid-scale point. It appears for some subjects that the meaning of "Fair" is distinctly favorable.

l. Acceptable.

"Acceptable" is not a desirable word to use as a mid-scale item. In one study it exhibited a marked bimodality of response, indicating that subjects disagreed on the degree of favorableness noted by the term. In a recent study, "Acceptable" had a large standard deviation of 1.46.

m Normal.

"Normal" is not a desirable word to use as a mid-scale item. In one study it exhibited a marked bimodality of response, indicating that the word "Normal" has different meanings for different subjects. This term would be classified as a synonym for "Average."

n. Medium.

"Medium" may possibly be used as a midpoint term. In one study there was a piling up of judgments for "Medium" at the middle scale position.

o. O.K. or all right.

"O.K." or "All right" have been used sometimes as midpoint response alternatives. However, they have a tendency to be rated somewhat positively. They also have larger standard deviations than other terms mentioned, indicating that there is ambiguity in their meaning.

p. So-so.

"So-so" is another term sometimes used as a midpoint response. In one study it had a scale value of 10.08, which was very close to the middle scale value of 10.00; but it also had a fairly large standard deviation of 1.87. Its use is not recommended.

q. Don't know.

"Don't know" is an unacceptable term to use as a midpoint. It usually means to the subject that, with additional knowledge or more time to think about the issue, he/she could choose one of the other alternatives.

r. Doesn't make any difference.

"Doesn't make any difference" should not be used as a midpoint response alternative because it implies a more negative value than a middle value. In one study it had a scale value of 2.83, where the middle scale value was 5.00. It also had an interquartile range of 3.13, which means that there was a lot of disagreement among subjects as to its meaning.

8. Selecting the Don't Know Response Alternative

Sometimes respondents are known for their tendency to mark the "Don't know" category. This selection is made when they are not aware of the content in a question or when they refuse to express their opinion. Researchers are not able to predict who would make a shift into a "Don't know" category when it is offered. It has been a common practice in public opinion survey research to leave out the "Don't know" category. Individuals who volunteer a "Don't know" response (even though it is not offered) will have it included as their selection of a response alternative. Human factors researchers who target surveys toward respondents who may not have access to specific experiences or equipment would appropriately include the "Don't know" category. When included in a survey, the "Don't know" category should be set apart from the other response alternatives to avoid confusing it with other categories. An example of the "Don't know" response alternatives is presented in Figure VIII-B-1.

Figure VIII-B-1

Inclusion of the "Don't Know" Response Alternative for a Maintenance Vehicle Questionnaire

Ease of Use Rating Scale

5	4	3	2	1	DK
Very Easy	Easy	Borderline	Difficult	Very Difficult	Don't Know

How easily can you:

1. "Gain access to the vehicle's batteries?	5	4	3	2	1	DK
2. Check battery and fluid levels?	5	4	3	2	1	DK
3. Check tightness of battery cables?"	5	4	3	2	1	DK

9. Selecting Positive and Negative Descriptors

If a balanced scale is desired, it is necessary to select an equal number of positive and negative descriptors. In most cases, it is easy to determine if a descriptor is positive or negative by seeing on which side of the (zero) midpoint its scale value falls. For example, "Mildly like" has a positive scale value, and "Mildly dislike" has a negative scale value.

Researchers at the Army Research Institute, Fort Hood, recommend avoiding the use of unbalanced directionality or intensity of attitude in the stem of a question. They usually work with rating scales similar to the semantic differential, which simplifies the composition of the stem. These researchers do not request a rating for how effective a system is, but instead they ask for a rating of how "effective-ineffective" the system is. Alternatively, they delete the dimension out of the stem altogether, and show the respondent the dimension only in the list of response alternatives. This approach is thought to create a formal balance in the response alternatives. Using these techniques, the stems either have a formal balance or avoid specifying the dimensionality of the rating.

The presentation of a positive or negative endpoint displayed first at the left-hand side of the scale has been investigated. It was found that order of presentation for the placement of positive or negative endpoints was not associated with response style (non-trait measures). Measures of personality traits are most influenced by balancing positive and negative descriptors or stems. Operational test and evaluation survey constructors would not need to be concerned about positioning the positive or negative endpoint first.

10. Selecting Type of Response Alternative

Points along the continuum of a scale have been anchored by many different types of response alternatives. For example, there have been response alternatives such as numbers, adjectives, adverbs, phrases, sentences, descriptions of behavior, etc. It does not seem to matter if response alternatives are numerical, verbal (one word), phrases, or behavioral. No one type of response alternative has proven superior to another.

11. Selecting Terms Showing Equal Intervals

Some experts argue that, in order to perform analyses on the basis of numerical values or weights, the intervals between rating scale response alternatives should be equal. This would be desirable, but in many cases it is impossible because many words have not been assigned scale values. But when scale values are available, the response alternatives can be selected as equally distant apart as possible when doing so is considered important.

There is a tendency for some questionnaire constructors to select phrases with parallel wording to indicate equal intervals. (They may also do so for other reasons.) However, if equal intervals are considered important, phrases should be selected based upon scale values if available. For example, in Table VIII-E-9, "Highly adequate" has a scale value of 3.843 while the parallel term "Highly inadequate" has a scale value of -4.196. This places "Highly inadequate" further away from the neutral point than "Highly adequate."

12. Use of Unscaled Terms

Some discussion is in order regarding the use of terms ignoring their scale values or to which no scale values have been assigned. An illustration of the first of these practices is from a study in which ARI had 21 Army officers involved in operational field testing rank-order 16 terms that included "Outstanding," "Superior," "Excellent" and "Very good." "Excellent" was ranked as less positive than "Outstanding" by 14 of the officers, while it was ranked as less positive than "Superior" by 17 of the officers. However, there was maximum disagreement as to whether "Outstanding" or "Superior" was first or second on the scale. That is, 12 rated "Superior" first and "Outstanding" second, while nine of the officers assigned the reverse ordering to these two words. All officers ranked "Outstanding," "Superior," and "Excellent" as more positive than "Very good." "Outstanding" is sometimes interpreted to denote only that the performance is among the best of a group -- without any implication as to quality, e.g., although a student's grade of 65 out of 100 points was failing, his/her performance may have been "Outstanding" since no other student in the class scored above 60!

What are the consequences to the developer of rating scale items of discovering a mean 50%-50% split as in the ordering of "Outstanding" and "Superior?" Does it mean they cannot be used together as part of the descriptors of a rating scale item? The answer is, "Normally yes." In Figure VIII-B-2, we would have better discrimination if "Outstanding" were replaced by "Excellent," with the position formerly occupied by "Excellent" being filled by "Very good." "Superior" and "Outstanding" or similarly overlapping terms should normally not be used on the same scale.

Figure VIII-B-2

Two Formats Using "Outstanding" and "Superior"

1. _____ 1. Superior

   _____ 2. Outstanding

   _____ 3. Excellent

   _____ 4. Good

   _____ 5. Fair

   _____ 6. Poor

2. Superior Outstanding Excellent Good Fair Poor

   (Circle one word)

When functioning as questionnaire consultants or developers in field test situations where respondents are enlisted personnel, ARI has recommended and used very little variety in its rating scale items. Arrays such as those shown in Figure VIII-B-3 are almost always proposed and used. Sometimes the middle term is deleted. Several reasons for the lack of variety are that a standard format facilitates: (1) comparability of rating distributions with previous tests, and (2) understanding by soldier respondents, who are often not high school graduates.

Figure VIII-B-3

Response Alternatives
Frequently Recommended by ARI

( ) Very satisfactory

( ) Satisfactory

( ) Borderline

( ) Unsatisfactory

( ) Very unsatisfactory

( ) Very effective

( ) Effective

( ) Borderline

( ) Ineffective

( ) Very ineffective

( ) Very acceptable

( ) Acceptable

( ) Borderline

( ) Unacceptable

( ) Very unacceptable

C. Selection of Response Alternatives Denoting Degrees of Frequency

Some questionnaire designers use verbal descriptors to denote degrees of frequency. Table VIII-C-1 shows such a list of verbal descriptors. A study showed that there was a great deal of variability in meaning for frequency phrases. Questionnaires should, whenever possible, use response alternatives that include a number designation or percentage of time meant by each word used as a response alternative.

Table VIII-C-1

Degrees of Frequency

Phrase	Scale Value	Inter-Quartile Range
Always	8.99	.52
Without fail	8.89	.61
Often	7.23	1.02
Usually	7.17	1.36
Frequently	6.92	.77
Now and then	4.79	1.40
Sometimes	4.78	1.83
Occasionally	4.13	2.06
Seldom	2.45	1.05
Rarely	2.08	.61
Never	1.00	.50

From: Dodd and Gerberick (1960).
See Section VIII-A 7.

D. Selection of Response Alternatives Using Order of Merit Lists of Descriptor Terms

An order of merit list of descriptors does not provide scale values nor show the variance of each phrase along some continuum. In addition, the list does not represent an equal interval scale. However, such lists are still useful for selecting response alternatives if the main concern is to select response categories so that each respondent will agree on the relative degree of "goodness" of the terms. Tables VIII-D-1 and VIII-D-2 give examples of order of merit lists of descriptor terms.

Table VIII-D-1

Order of Merit of Selected Descriptive Terms

Order of Merit	Descriptive Term
1	Very superior
2	Very outstanding
3	Superior
4	Outstanding
5	Excellent
6	Very good
7	Good
8	Very satisfactory
9	Satisfactory
10	Marginal
11	Borderline
12	Poor
13	Unsatisfactory
14	Bad
15	Very poor
16	Very unsatisfactory
17	Very bad
18	Extremely poor
19	Extremely unsatisfactory
20	Extremely bad

From: Gividen (1973). Section VIII-A 9.

Table VIII-D-2

Order of Merit of Descriptive Terms
Using "Use" as a Descriptor

Order of Merit	Descriptive Term
1	Extremely useful
2	Very useful
3	Of significant use
4	Of considerable use
5	Of much use
6	Of moderate use
7	Of use
8	Of some use
9	Of little use
10	Not very useful
11	Of slight use
12	Of very little use
13	Of no use

From: Givlden (1973). See Section VIII-A 9.

E. Selection of Response Alternatives Using Scale Values and Standard Deviations

Using scale values and standard deviations to select response alternatives will give a more refined set of phrases than using an order of merit list. Other sections above have discussed specific considerations in selecting descriptors. In general, response alternatives selected from lists of phrases with scale values should usually have the following characteristics:

1. The scale values of the terms should be as far apart as possible.

2. The scale values of the terms should be as equally distant as possible.

3. The terms should have small variability (small standard deviations or interquartile ranges).

4. Other things being equal, the terms should have parallel wording.

Tables VIII-E-1 through VIII-E-24 give lists of phrases which have scale values and, when possible, standard deviations or interquartile range. They are based on empirical evidence, and may be used to select response alternatives.

Table VIII-E-1

Acceptability Phrases

Phrase	Average	SD
Excellent	6.27	.54
Perfect in every respect	6.22	.86
Extremely good	5.74	.81
Very good	5.19	.75
Unusually good	5.03	.98
Very good in most respects	4.62	.72
Good	4.25	.90
Moderately good	3.58	.77
Could use some minor changes	3.28	1.09
Not good enough for extreme conditions	3.10	1.30
Not good for rough use	2.72	1.15
Not very good	2.10	.85
Needs major changes	1.97	1.12
Barely acceptable	1.79	.90
Not good enough for general use	1.76	1.21
Better than nothing	1.22	1.08
Poor	1.06	1.11
Very poor	.76	.95
Extremely poor	.36	.76

From: U.S. Army (1973). See Section VIII-A 21.

Table VIII-E-2

Degrees of Excellence: First Set

Phrase	Scale Value	SD
Superior	20.12	1.17
Fantastic	20.12	0.83
Tremendous	19.84	1.31
Superb	19.80	1.19
Excellent	19.40	1.73
Terrific	19.00	2.45
Outstanding	18.96	1.99
Wonderful	17.32	2.30
Delightful	16.92	1.85
Fine	14.80	2.12
Good	14.32	2.08
Pleasant	13.44	2.06
Nice	12.56	2.14
Acceptable	11.12	2.59
Average	10.84	1.55
All right	10.76	1.42
O.K.	10.28	1.67
Neutral	9.80	1.50
Fair	9.52	2.06
Mediocre	9.44	1.80
Unpleasant	5.04	2.82
Bad	3.88	2.19
Very bad	3.20	2.10
Unacceptable	2.64	2.04
Awful	1.92	1.50
Terrible	1.76	.77
Horrible	1.48	.87

From: Myers and Warner (1968).
See Section VIII-A 17.

Table VIII-E-3

Degrees of Excellence: Second Set

Phrase	Scale Value	SD
Best of all	6.15	2.48
Excellent	3.71	1.01
Wonderful	3.51	.97
Mighty fine	2.88	.67
Especially good	2.86	.82
Very good	2.56	.87
Good	1.91	.76
Pleasing	1.58	.65
O.K.	.87	1.24
Fair	.78	.85
Only fair	.71	.64
Not pleasing	-.83	.67
Poor	-1.55	.87
Bad	-2.02	.80
Very bad	-2.53	.64
Terrible	-3.09	.98

From: Jones and Thurstone (1955).
See Section VIII-A 12.

Table VIII-E-4

Degrees of Like and Dislike

Phrase	Scale Value	SD
Like extremely	4.16	1.62
Like intensely	4.05	1.59
Strongly like	2.96	.69
Like very much	2.91	.60
Like very well	2.60	.78
Like quite a bit	2.32	.52
Like fairly well	1.51	.59
Like	1.35	.77
Like moderately	1.12	.61
Mildly like	.85	.47
Like slightly	.69	.32
Neutral	.02	.18
Like not so well	-.30	1.07
Like not so much	-.41	.94
Dislike slightly	-.59	.27
Mildly dislike	-.74	.35
Dislike moderately	-1.20	.41
Dislike	-1.58	.94
Don't like	-1.81	.97
Strongly dislike	-2.37	.53
Dislike very much	-2.49	.64
Dislike intensely	-3.33	1.39
Dislike extremely	-4.32	1.86

From: Jones and Thurstone (1955).
See Section VIII-A 12.

Table VIII-E-5

Degrees of Good and Poor

Phrase	Scale Value	SD
Exceptionally good	18.56	2.36
Extremely good	18.44	1.61
Unusually good	17.08	2.43
Remarkably good	16.68	2.19
Very good	15.44	2.77
Quite good	14.44	2.76
Good	14.32	2.08
Moderately good	13.44	2.23
Reasonably good	12.92	2.93
Fairly good	11.96	2.42
Slightly good	11.84	2.19
So-so	10.08	1.87
Not very good	6.72	2.82
Moderately poor	6.44	1.64
Reasonably poor	6.32	2.46
Slightly poor	5.92	1.96
Poor	5.72	2.09
Fairly poor	5.64	1.68
Quite poor	4.80	1.44
Unusually poor	3.20	1.44
Very poor	3.12	1.17
Remarkably poor	2.88	1.74
Exceptionally poor	2.52	1.19
Extremely poor	2.08	1.19

From: Myers and Warner (1968).
See Section VIII-A 17.

Table VIII-E-6

Degrees of Good and Bad

Phrase	Scale Value
Extremely good	3.449
Very good	3.250
Unusually good	3.243
Decidedly good	3.024
Quite good	2.880
Rather good	2.755
Good	2.712
Pretty good	2.622
Somewhat good	2.462
Slightly good	2.417
Slightly bad	1.497
Somewhat bad	1.323
Rather bad	1.232
Bad	1.024
Pretty bad	1.018
Quite bad	.924
Decidedly bad	.797
Unusually bad	.662
Very bad	.639
Extremely bad	.470

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-7

Degrees of Agree and Disagree

Phrase	Mean	SD
Decidedly agree	2.77	.41
Quite agree	2.37	.49
Considerably agree	2.21	.42
Substantially agree	2.10	.50
Moderately agree	1.47	.41
Somewhat agree	.94	.41
Slightly agree	.67	.36
Perhaps agree	.52	.46
Perhaps disagree	-.43	.46
Slightly disagree	-.64	.38
Somewhat disagree	-.93	.47
Moderately disagree	-1.35	.42
Quite disagree	-2.16	.57
Substantially disagree	-2.17	.51
Considerably disagree	-2.17	.45
Decidedly disagree	-2.76	.43

From: Altemeyer (1970).
See Section VIII-A 1.

Table VIII-E-8

Degrees of More and Less

Phrase	Scale Value	Inter-quartile Range
Very much more	8.02	.61
Much more	7.67	1.04
A lot more	7.50	1.06
A good deal more	7.29	.98
More	6.33	1.01
Somewhat more	6.25	.98
A little more	6.00	.58
Slightly more	5.99	.57
Slightly less	3.97	.56
A little less	3.96	.54
Less	3.64	1.04
Much less	2.55	1.06
A good deal less	2.44	1.11
A lot less	2.36	1.03
Very much less	1.96	.52

From: Dodd and Gerberick (1960).
See Section VIII-A 7.

Table VIII-E-9

Degrees of Adequate and Inadequate

Phrase	Mean	SD
Totally adequate	4.620	.846
Absolutely adequate	4.540	.921
Completely adequate	4.490	.825
Extremely adequate	4.412	.719
Exceptionally adequate	4.380	.869
Entirely adequate	4.340	.863
Wholly adequate	4.314	1.038
Fully adequate	4.294	.914
Very very adequate	4.063	.876
Perfectly adequate	3.922	1.026
Highly adequate	3.843	.606
Most adequate	3.843	.978
Very adequate	3.420	.851
Decidedly adequate	3.140	1.536
Considerably adequate	3.020	.874
Quite adequate	2.980	.979
Largely adequate	2.863	.991
Substantially adequate	2.608	1.030
Reasonably adequate	2.412	.771
Pretty adequate	2.306	.862
Rather adequate	1.755	.893
Mildly adequate	1.571	.670
Somewhat adequate	1.327	.793
Slightly adequate	1.200	.566
Barely adequate	.627	.928
Neutral	.000	.000
Borderline	-.020	.316
Barely inadequate	-1.157	.638
Mildly inadequate	-1.353	.621
Slightly inadequate	-1.380	.772
Somewhat inadequate	-1.882	.732
Rather inadequate	-2.102	.974
Moderately inadequate	-2.157	1.017
Fairly inadequate	-2.216	.800
Pretty inadequate	-2.347	.959
Considerably inadequate	-3.600	.680
Very inadequate	-3.735	.777
Decidedly inadequate	-3.780	.944
Most inadequate	-3.980	1.545
Highly inadequate	-4.196	.741

(Table continued on next page)

Table VIII-E-9 (Cont.)

Degrees of Adequate and Inadequate

Phrase	Mean	SD
Very very inadequate	-4.460	.537
Extremely inadequate	-4.608	.527
Fully inadequate	-4.667	.676
Exceptionally inadequate	-4.680	.508
Wholly inadequate	-4.784	.498
Entirely inadequate	-4.792	.644
Completely inadequate	-4.800	.529
Absolutely inadequate	-4.880	.431
Totally inadequate	-4.900	.412

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-E-10

Degrees of Acceptable and Unacceptable

Phrase	Mean	SD
Wholly acceptable	4.725	.563
Completely acceptable	4.686	.610
Fully acceptable	4.412	.867
Extremely acceptable	4.392	.716
Most acceptable	4.157	.915
Very very acceptable	4.157	.825
Highly acceptable	4.040	.631
Quite acceptable	3.216	.956
Largely acceptable	3.137	.991
Acceptable	2.392	1.456
Reasonably acceptable	2.294	.722
Moderately acceptable	2.280	.722
Pretty acceptable	2.000	1.125

(Table continued on next page)

Table VIII-E-10 (Cont.)

Degrees of Acceptable and Unacceptable

Phrase	Mean	SD
Rather acceptable	1.939	.818
Fairly acceptable	1.840	.924
Mildly acceptable	1.686	.700
Somewhat acceptable	1.458	1.241
Barely acceptable	1.078	.518
Slightly acceptable	1.039	.522
Sort of acceptable	.940	.645
Borderline	.000	.200
Neutral	.000	.000
Marginal	-.120	.515
Barely unacceptable	-1.100	.300
Slightly unacceptable	-1.255	.589
Somewhat unacceptable	-1.765	.674
Rather unacceptable	-2.020	.836
Fairly unacceptable	-2.160	.880
Moderately unacceptable	-2.340	.681
Pretty unacceptable	-2.412	.662
Reasonably unacceptable	-2.440	.753
Unacceptable	-2.667	1.381
Substantially unacceptable	-3.235	.899
Quite unacceptable	-3.388	1.066
Largely unacceptable	-3.392	.818
Considerably unacceptable	-3.440	.779
Notably unacceptable	-3.500	1.044
Decidedly unacceptable	-3.837	1.017
Highly unacceptable	-4.294	.535
Most unacceptable	-4.420	.724
Very very unacceptable	-4.490	.500
Exceptionally unacceptable	-4.540	.607
Extremely unacceptable	-4.686	.464
Completely unacceptable	-4.900	.361
Entirely unacceptable	-4.900	.361
Wholly unacceptable	-4.922	.269
Absolutely unacceptable	-4.922	.334
Totally unacceptable	-4.941	.235

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-E-11

Comparison Phrases

Phrase	Mean	SD
Best of all	4.896	.510
Absolutely best	4.843	.459
Truly best	4.600	.721
Undoubtedly best	4.569	.823
Decidedly best	4.373	.839
Best	4.216	1.459
Absolutely better	4.060	.988
Extremely better	3.922	.882
Substantially best	3.700	.922
Decidedly better	3.412	.933
Conspicuously better	3.059	.802
Moderately better	2.255	.737
Somewhat better	1.843	.801
Rather better	1.816	.719
Slightly better	1.157	.776
Barely better	.961	.656
Absolutely alike	.588	1.623
Alike	.216	.847
The same	.157	.801
Neutral	.000	.000
Borderline	-.061	.314
Marginal	-.184	.919
Barely worse	-1.039	.816
Slightly worse	-1.216	.498
Somewhat worse	-2.078	.860
Moderately worse	-2.220	.944
Noticeably worse	-2.529	1.036
Worse	-2.667	1.423
Notably worse	-3.020	1.038
Largely worse	-3.216	1.108
Considerably worse	-3.275	1.206
Conspicuously worse	-3.275	.887
Much worse	-3.286	.808
Substantially worse	-3.460	.899
Decidedly worse	-3.760	.907
Very much worse	-3.941	.752
Absolutely worse	-4.431	.823
Decidedly worst	-4.431	.748
Undoubtedly worst	-4.510	.872
Absolutely worst	-4.686	1.291
Worst of all	-4.776	1.298

From: Matthews, Wright, and Yudowitch (1975).
See Section VIII-A 13.

Table VIII-E-12

Degrees of Satisfactory and Unsatisfactory

Phrase	Scale Value	SD
Quite satisfactory	4.35	.95
Satisfactory	3.69	.87
Not very satisfactory	2.11	.76
Unsatisfactory but usable	2.00	.87
Very unsatisfactory	.69	1.32

From: U.S. Army (1973). See Section VIII-A 21.

Table VIII-E-13

Degrees of Unsatisfactory

Phrase	Scale Value
Unsatisfactory	1.47
Quite unsatisfactory	1.00
Very unsatisfactory	.75
Unusually unsatisfactory	.75
Highly unsatisfactory	.71
Very, very unsatisfactory	.25
Extremely unsatisfactory	.10
Completely unsatisfactory	.00

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-14

Degrees of Pleasant

Phrase	Scale Value
Extremely pleasant	3.490
Very pleasant	3.174
Unusually pleasant	3.107
Decidedly pleasant	3.028
Quite pleasant	2.849
Pleasant	2.770
Rather pleasant	2.743
Pretty pleasant	2.738
Somewhat pleasant	2.505
Slightly pleasant	2.440

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-15

Degrees of Agreeable

Phrase	Scale Value
Very, very agreeable	5.34
Extremely agreeable	5.10
Highly agreeable	5.02
Completely agreeable	4.96
Unusually agreeable	4.86
Very agreeable	4.82
Quite agreeable	4.45
Agreeable	4.19

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-16

Degrees of Desirable

Phrase	Scale Value
Very, very desirable	5.66
Extremely desirable	5.42
Completely desirable	5.38
Unusually desirable	5.23
Highly desirable	5.15
Very desirable	4.96
Quite desirable	4.76
Desirable	4.50

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-17

Degrees of Nice

Phrase	Scale Value
Extremely nice	3.351
Unusually nice	3.155
Very nice	3.016
Decidedly nice	2.969
Pretty nice	2.767
Quite nice	2.738
Nice	2.636
Rather nice	2.568
Somewhat nice	2.488
Slightly nice	2.286

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-18

Degrees of Adequate

Phrase	Scale Value	SD
More than adequate	4.13	1.11
Adequate	3.39	.87
Not quite adequate	2.40	.85
Barely adequate	2.10	.84
Not adequate	1.83	.98

From: U.S. Army (1973).
See Section VIII-A 21.

Table VIII-E-19

Degrees of Ordinary

Phrase	Scale Value
Ordinary	2.074
Very ordinary	2.073
Somewhat ordinary	2.038
Rather ordinary	2.034
Pretty ordinary	2.026
Slightly ordinary	1.980
Decidedly ordinary	1.949
Extremely ordinary	1.936
Unusually ordinary	1.875

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-20

Degrees of Average

Phrase	Scale Value
Rather average	2.172
Average	2.145
Quite average	2.101
Pretty average	2.094
Somewhat average	2.080
Unusually average	2.062
Extremely average	2.052
Very average	2.039
Slightly average	2.023
Decidedly average	2.020

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-21

Degrees of Hesitation

Phrase	Scale Value	Inter-quartile Range
Without hesitation	7.50	6.54
With little hesitation	5.83	3.40
Hesitant	4.77	1.06
With some hesitation	4.38	1.60
With considerable hesitation	3.29	3.39
With much hesitation	3.20	5.25
With great hesitation	2.41	6.00

From: Dodd and Gerberick (1960). See Section VIII-A 7.

Table VIII-E-22

Degrees of Inferior

Phrase	Scale Value
Slightly inferior	1.520
Somewhat inferior	1.516
Inferior	1.323
Rather inferior	1.295
Pretty inferior	1.180
Quite inferior	1.127
Decidedly inferior	1.013
Unusually inferior	.963
Very inferior	.927
Extremely inferior	.705

From: Cliff (1959).
See Section VIII-A 6.

Table VIII-E-23

Degrees of Poor

Phrase	Scale Value
Poor	1.60
Quite poor	1.30
Very poor	1.18
Unusually poor	.95
Extremely poor	.95
Completely poor	.92
Very, very poor	.55

From: Mosier (1941).
See Section VIII-A 16.

Table VIII-E-24

Descriptive Phrases

Phrase	Scale Value	Inter-quartile Range
Complete	8.85	.65
Extremely vital	8.79	.84
Very certain	8.55	1.05
Very strongly	8.40	1.04
Very crucial	8.29	1.12
Very important	8.22	1.16
Very sure	8.15	.95
Almost complete	8.06	.58
Of great importance	8.05	.91
Very urgent	8.00	.90
Feel strongly toward	7.80	1.60
Essential	7.58	1.85
Very vital	7.55	1.05
Certain	7.13	1.44
Strongly	7.07	.67
Important	6.83	1.14
Good	6.72	1.20
Urgent	6.41	1.53
Crucial	6.39	1.73
Sure	5.93	1.87
Vital	5.92	1.63
Moderately	5.24	.99
Now	5.03	.53
As at present	5.00	.50
Fair	4.96	.77
Don't know	4.82	.82
Undecided	4.76	1.06
Don't care	4.63	2.00
Somewhat	3.79	.94
Indifferent	3.70	2.20
Object strongly to	3.50	6.07
Not important	3.09	1.33
Unimportant	1.94	1.42
Bad	2.83	.93
Uncertain	2.83	2.50
Doesn't make any difference	2.83	3.13
Not sure	2.82	1.24
Not certain	2.64	2.62

(Table continued on next page)

Table VIII-E-24 (Cont.)

Descriptive Phrases

Phrase	Scale Value	Inter-quartile Range
Non-essential	2.58	1.67
Doesn't mean anything	2.50	2.71
Insignificant	2.12	1.14
Very little	2.08	.64
Almost none	2.04	.57
Very unimportant	1.75	1.25
Only as a last resort	1.70	7.30
Very bad	1.50	1.13
None	1.11	.59

From: Dodd and Gerberick (1960).
See Section VIII-A 7.

F. Sample Sets of Response Alternatives

It is sometimes valuable and is a time-saver to have lists of response alternatives available to use. The tables in this section give some examples of response alternatives that have been selected on different bases. These sets do not exhaust all possibilities.

The sets of response alternatives that appear in Table VIII-F-1 were selected so that the phrases in each set would have means at least one standard deviation away from each other and have parallel wording. Some of the sets of response alternatives have extreme endpoints, some do not. The sets of response alternatives shown in Table VIII-F-2 were selected so that the phrases in each set would be as nearly equally distant from each other as possible without regard to parallel wording. Table VIII-F-3 contains sets of response alternatives selected from lists of descriptors with only scale values given. The phrases were selected on the bases of equal appearing intervals. Table VIII-F-4 has sets of response alternatives selected from order of merit lists of descriptors.

Table VIII-F-1

Sets of Response Alternatives Selected So Phrases Are at Least One Standard Deviation Apart and Have Parallel Wording

Set No.	Response Alternatives	Set No.	Response Alternatives
1.	Completely acceptable	7.	Very adequate
	Reasonably acceptable		Slightly adequate
	Barely acceptable		Borderline
	Borderline		Slightly inadequate
	Barely unacceptable		Very inadequate
	Reasonably unacceptable		
	Completely unacceptable	8.	Highly adequate
			Mildly adequate
2.	Wholly acceptable		Borderline
	Largely acceptable		Mildly inadequate
	Borderline		Highly inadequate
	Largely unacceptable		
	Wholly unacceptable	9.	Decidedly agree
			Substantially agree
3.	Largely acceptable		Slightly agree
	Barely acceptable		Slightly disagree
	Borderline		Substantially disagree
	Barely unacceptable		Decidedly disagree
	Largely unacceptable		
		10.	Moderately agree
4.	Reasonably acceptable		Perhaps agree
	Slightly acceptable		Neutral
	Borderline		Perhaps disagree
	Slightly unacceptable		Moderately disagree
	Reasonably unacceptable		
		11.	Undoubtedly best
5.	Totally adequate		Conspicuously better
	Very adequate		Moderately better
	Barely adequate		Alike
	Borderline		Moderately worse
	Barely inadequate		Conspicuously worse
	Very inadequate		Undoubtedly worst
	Totally inadequate		
		12..	Moderately better
6.	Completely adequate		Barely better
	Considerably adequate		The same
	Borderline		Barely worse
	Considerably inadequate		Moderately worse
	Completely inadequate		

(Table continued on next page)

Table VIII-F-1 (Cont.)

Sets of Response Alternatives Selected So Phrases Are at Least One Standard Deviation Apart and Have Parallel Wording

Set No.	Response Alternatives	Set No.	Response Alternatives
13.	Extremely good	16.	Like extremely
	Remarkably good		Like moderately
	Good		Neutral
	So-so		Dislike moderately
	Poor		Dislike extremely
	Remarkably poor		
	Extremely poor	17.	Strongly like
			Like
14.	Exceptionally good		Neutral
	Reasonably good		Don't like
	So-so		Strongly dislike
	Reasonably poor		
	Exceptionally poor	18.	Very much more
			A good deal more
15.	Very important		A little more
	Important		A little less
	Not important		A good deal less
	Very unimportant		Very much less

Table VIII-F-2

Sets of Response Alternatives Selected So That
Intervals Between Phrases Are as Nearly Equal as Possible

Set No.	Response Alternatives	Set No.	Response Alternatives
1.	Completely acceptable	7.	Perfect in every respect
	Reasonably acceptable		Very good
	Borderline		Good
	Moderately unacceptable		Could use some minor changes
	Extremely unacceptable		Not very good
			Better than nothing
2.	Totally adequate		Extremely poor
	Pretty adequate		
	Borderline	8.	Excellent
	Pretty inadequate		Good
	Extremely inadequate		Only fair
			Poor
3.	Highly adequate		Terrible
	Rather adequate		
	Borderline	9.	Extremely good
	Somewhat inadequate		Quite good
	Decidedly inadequate		So-so
			Slightly poor
4.	Quite agree		Extremely poor
	Moderately agree		
	Perhaps agree	10.	Remarkably good
	Perhaps disagree		Moderately good
	Moderately disagree		So-so
	Substantially disagree		Not very good
			Unusually poor
5.	Undoubtedly best		
	Moderately better	11.	Without hesitation
	Borderline		With little hesitation
	Noticeably worse		With some hesitation
	Undoubtedly worst		With great hesitation
6.	Fantastic	12.	Strongly like
	Delightful		Like quite a bit
	Nice		Like
	Mediocre		Neutral
	Unpleasant		Mildly dislike
	Horrible		Dislike very much
			Dislike extremely

(Table continued on next page)

Table VIII-F-2 (Cont.)

Sets of Response Alternatives Selected So That
Intervals Between Phrases Are as Nearly Equal as Possible

Set No.	Response Alternatives	Set No.	Response Alternatives
13.	Like quite a bit Like Like slightly Borderline Dislike slightly Dislike moderately Don't like	15.	Very much more A little more Slightly less Very much less
14.	Like quite a bit Like fairly well Borderline Dislike moderately Dislike very much		

Table VIII-F-3

Sets of Response Alternatives Selected
from Lists Giving Scale Values Only

Set No.	Response Alternatives	Set No.	Response Alternatives
1.	Very, very agreeable	6.	Extremely nice
	Usually agreeable		Decidedly nice
	Quite agreeable		Nice
	Agreeable		Slightly nice
2.	Rather average	7.	Ordinary
	Quite average		Slightly ordinary
	Unusually average		Unusually ordinary
	Decidedly average		
3.	Very, very desirable	8.	Extremely pleasant
	Completely desirable		Decidedly pleasant
	Very desirable		Somewhat pleasant
	Desirable		
4.	Extremely good	9.	Poor
	Somewhat good		Very poor
	Slightly bad		Very, very poor
	Extremely bad		
5.	Slightly inferior	10.	Very, very agreeable
	Rather inferior		Extremely agreeable
	Unusually inferior		Very agreeable
	Extremely inferior		Quite agreeable
			Agreeable

Note. Selected so that intervals between phrases are as equal as possible.

Table VIII-F-4

Sets of Response Alternatives Selected
Using Order of Merit Lists of Descriptor Terms

Set No.	Response Alternatives
1.	Very good
	Good
	Borderline
	Poor
	Very poor
2.	Very satisfactory
	Satisfactory
	Borderline
	Unsatisfactory
	Very unsatisfactory
3.	Very superior
	Superior
	Borderline
	Poor
	Very poor
4.	Extremely useful
	Of considerable use
	Of use
	Not very useful
	Of no use

Chapter IX: Physical Characteristics of Questionnaires

A. Overview

This chapter considers five topics related to the physical characteristics of questionnaires: the location of response alternatives relative to the stem (Section IX-B); questionnaire length (Section IX-C); questionnaire format considerations (Section IX-D); the use of answer sheets (Section IX-E); and the use of branching (Section IX-E).

B. Location of Response Alternatives Relative to the Stem

Research to determine what effect the location of response alternatives relative to the question stem has on subjects' responses is practically nonexistent. There is some evidence, however, that untrained raters can make relatively error-free graphic ratings regardless of whether the "good" end of the scale is at the left, right, top, or bottom.

In designing a specific questionnaire, the following points should be considered regarding the location of response alternatives relative to the stem:

1. With multiple choice items, the response alternatives are usually arranged vertically under the stem as shown in Figure IV-C-1. With a large number of response alternatives, two or more columns of vertically arranged alternatives might be used. Sometimes, if there are only two or three alternatives (such as "Yes" and "No"), they are placed horizontally rather than vertically.

2. Graphic rating scales are usually placed horizontally on a page. However, the descriptive words, phrases, or sentences on a scale should be concentrated as much as possible at specific points on the scale. This is usually easier if the scales are placed vertically on the page, but it can be done either way. Descriptors need not be equally spaced along graphic scales, and should not be if there is reason to believe the psychological distances between them are not equal.

3. With nongraphic (or "numerical") rating scale items and with ranking and forced choice items, the response alternatives are usually placed vertically under the question stem. See examples in Chapter IV. Sometimes rating scale items are placed horizontally under the stem as shown in Figure VII-B-1. If a number of rating scale items all use the same response alternatives, the question stems can be presented in a column with the response alternatives to the right as shown in Figure IX-B-1.

   In Figure IX-B-1, the response alternatives have been rotated 90 degrees to save space. An effort should be made to place the response alternative horizontal with the bottom of the page so that the respondent does not need to turn the page sideways to read them.

4. The response alternatives for semantic differential items are usually placed horizontally on the page. For an example, see Figure IV-I-1.

5. Use precoded cards with alphabet letters for responses to items with sensitive content that might be viewed as threatening. This would be appropriate for questionnaires administered by personal interview. Selection by respondent of the alphabet letter can later be transposed to numbers for analysis purposes. This technique is used to obtain less distortion to reduce the social desirability response set.

Figure IX-B-1

Arrangement of Items With Same
Rating Scale Response Alternatives

1. How satisfied or dissatisfied are you with each of the following factors or things?

	Very Satisfied	Satisfied	Borderline	Dissatisfied	Very Dissatisfied
a. Type of furniture in barracks.	___	___	___	___	___
b. Medical service to soldiers.	___	___	___	___	___
c. Quality of mess hall food.	___	___	___	___	___
d. Leadership of generals.	___	___	___	___	___
e. Opportunity for promotion.	___	___	___	___	___
f. Army pay.	___	___	___	___	___
g. Civilian opinion of Army.	___	___	___	___	___

6. For respondents with a low education level, an easy format with stems and anchors easy to understand is essential. Sometimes, respondents have a preference for questionnaire formats with which they have had previous experience. However, preference for specific kinds of formats does not mean that the results will be more reliable. In some studies, respondents were more accurate in their ratings with less preferred formats.

C. Questionnaire Length

1. General

The length of questionnaires used in field tests has ranged from one page to as many as 30 pages, perhaps more. How long can one expect a respondent to work effectively at the questionnaire-answering task? At what point do attention and motivation start to degrade, thereby producing poorly considered responses or the omission of responses? Research information on this point is not available to provide a basis for a firm recommendation. There is even disagreement on the effect of questionnaire length on the response rate to mailed questionnaires. The number of items and number of pages in a questionnaire may not necessarily be related to response rate for mailed questionnaires.

However, questionnaires which require longer than one hour to complete will, in most situations, cause boredom and indifference. Even 10 or 15 minutes may be too long if the questionnaire is perceived by the respondent as redundant or asking unnecessary questions. If one is concerned about the effects of a long questionnaire, alternate forms should be used, wherein the order of items is reversed (or approximately so). For example, the items answered last on 50% of the forms would be answered first on the other 50% of the forms. One could also split the respondent group in half and give half of the questions to each group--provided that the two groups were fairly equivalent in relevant characteristics. Splitting the respondent group in half increases the complexity of the survey and may affect the precision of the measures. It is assumed that everything else would already have been done to reduce the number of items before one of these approaches is used.

For questionnaires administered by interview, survey guidelines were established by the Federal Office of Management and Budget. They suggest that interviews should not take longer than half an hour, although there may be valid reasons why more information would be required. This would, of course, extend the interview length. Many surveys take an hour or longer to complete. There are no firm guidelines. Pretesting the questionnaire may provide data on the effect of number of items on the response rate.

2. Results of a Study

In a 1976 study, ARI assisted TCATA in obtaining and analyzing questionnaire responses from a group of trainees whose duration and location of basic and advanced individual training was handled differently from the usual. The number of trainees answering items 1-7 and 48-54 of a 54-item questionnaire is shown below. Note that there is very little drop in the number of men in either group as we skip from items 1-7 to items 48-52. This suggests that a 50-item questionnaire, administered as this was, was not so long that persons stopped responding after answering successively more questions.

Now note the sharp drop-about 15% and 9% for the two groups-in responses to items 53 and 54. A more gradual decrease in number of people responding is more what one would expect if they are being "worn down" or fatigued by excessive length.

This result was puzzling, but then it was noted that items 53 and 54 are alone together on the tenth and final page of the questionnaire. It is speculated that many/most of those not answering items 53 and 54 turned page 10 over along with page 9 and thought they had answered all that was required of them. No one checked their questionnaires when they were handed in to see if they had left any items blank. The reductions in respondents appears more of a "last page phenomena" than a consequence of an excessively long questionnaire.

Item #	Experimental Group	Control Group
1	716	512
2	716	513
3	717	511
4	714	513
5	716	514
6	713	510
7	716	511
:	:	:
48	707	509
49	707	508
50	707	508
51	707	510
52	698	505
53	593	462
54	604	461

D. Questionnaire Format Considerations

This section addresses the format of questionnaire items, title and other identification marks, printed introductions, planning to facilitate processing, and other questionnaire format considerations.

1. Format of Questionnaire Items and Format Bias

Item format biases occur when responses to items (questions) are influenced by the question stem or response alternatives. The following guidance is provided:

a. The format of all questionnaire items on a questionnaire should be consistent whenever possible. Mixing multiple choice questions, open-ended questions, scales, etc. is normally not desirable.

b. Punctuation and question structure should be consistent and in accordance with proper sentence structure principles. Where incomplete sentences (e.g., "The training that I have received at Fort Hood has been" with five response alternatives of "very challenging" through "very unchallenging") are used as stems, no extraneous punctuation, such as a colon, need be put at the end of the stem. The first word of the response alternatives should not be capitalized unless they would be if the statement were written as a continuous sentence. Terminal punctuation at the end of the response alternatives should follow the same general rule of consistency with normal sentence structure. Hence, a period would ordinarily be placed after each response alternative.

When an item consists of a complete question (e.g., "How satisfied or dissatisfied are you with the furniture in the barracks?"), the first word of the response alternatives should be capitalized since it does not continue a sentence. If the response alternatives constitute complete sentences, then they should have periods at the end, or whatever other terminal punctuation is appropriate. Sometimes periods are placed at the end of extremely long response alternatives even if they are not sentences. Ordinarily, then, with this form of item, periods would not be placed after the response alternatives.

Exceptions to the above suggestions should be made whenever the exception would improve clarity. An example might be when periods would be confused with decimal points.

c. When items are ambiguous, a recognizable pattern of inappropriate responses is often produced.

d. Item format bias may be a function of how items are sequenced and grouped.

e. Some authors conclude that a bias can be expected from all closed-end questions where answers must be selected from two or more fixed choices.

f. The paired-comparison format may be useful for those respondents who tend to check many items from a list, and for those who check only a few.

g. Card sorting may show the least item format bias.

h. With two-way choices, some respondents have a tendency to select the first alternative. Others have a tendency to select the second. With other multiple choice items, some respondents have a tendency to select certain categories.

i. There is some evidence that the first response alternative to a question is chosen somewhat more frequently than the others.

j. Two studies were conducted by Mayer and Piper (1982) regarding physical layout of a questionnaire. Questionnaire layout can be confusing to respondents. The wrong categories were initially marked by respondents indicating erroneous brand preferences. An example is provided in Figure IX-D-1 to illustrate how modification of the questionnaire format facilitated clarification. The questionnaire layout that confused respondents did not have a response alternative for "Other brand." The layout was identical to that of Brand A through Brand G response alternatives. There was no bracketed response alternative for "Other brand."

Mayer, C. S., & Piper, C. (1982). A note on the importance of layout in self-administered questionnaires. Journal of Marketing Research, 19(3), 390-391.

Figure IX-D-1

Original Questionnaire Format

	Product X	Product Y	Product Z
Brand F ---	( )6	( )6	( )6
Brand G ---	( )7	( )7	( )7
Other brand (SPECIFY)	_____	_____	_____

Modified Questionnaire Format

	Product X	Product Y	Product Z
Brand F ---	( )6	( )6	( )6
Brand G ---	( )7	( )7	( )7
Other brand ___	( )8	( )8	( )8

2. Title and Other Identification Marks

Each questionnaire should carry a descriptive title centered at the top of the first page of questions and on the instructional and/or introductory cover page if such is used. Each questionnaire form should also be designated by form number to distinguish it from other forms. This number usually goes in the upper left-hand corner of each page.

3. Printed Introductions

Introductions are sometimes printed at the start of a questionnaire to tell respondents the purpose and importance of the questionnaire, and the importance of their cooperation in answering all questions carefully. Methodological research is needed to determine the effectiveness of such introductions, but if they are too lengthy, there is always the possibility that they might be counterproductive. Regardless, if the introduction is going to run more than a quarter of a page, it might better be placed on a cover sheet.

See Section X-B regarding the content of questionnaire instructions.

4. Planning to Facilitate Processing

Where possible, questionnaires should be designed to facilitate data collection, reduction, and analysis. This frequently involves formulating the questionnaire for machine processing. For small samples, however, manual processing should normally be employed since the effort needed to plan for machine processing is not justified by anticipated data reduction time savings. How to format a questionnaire for machine processing is outside the current scope of this manual. See Section IX-E regarding the use of answer sheets.

5. Other Questionnaire Format Considerations

a. If the respondent's name, rank, etc. is really needed, ask for it on the front page. (See also Section X-C.) Sometimes other information is needed about respondents so that it can be correlated with their responses. This may include duty MOS, special army training, combat experience, etc. If it is really needed, it is usually asked for on the front page along with name.

b. If a questionnaire has over two pages, page numbers should be used. They are ordinarily put at the center bottom of each page.

c. A questionnaire should not be crowded or cluttered in appearance. If it is, certain items might be missed.

d. Each item in a questionnaire should be numbered or lettered so it can be identified and referred to.

e. Sufficient room should be left for the respondent to write in answers to open-ended questions.

f. Directions should be well displayed and unmistakably clear.

g. There is research evidence that an attractive questionnaire increases response rates.

h. Different colored pages or questionnaire forms may aid in the sorting of data and may have appeal to the respondents.

E. Use of Answer Sheets

As noted in Section IX-D 4, when possible, questionnaires should be designed to facilitate data collection, processing, and analyses. Hence, if the number of questions warrant it, consideration should be given to the use of separate answer sheets. An answer sheet can be designed for either hand or machine processing.

When considering the possible use of answer sheets, the following points should be kept in mind:

1. The use of a separate answer sheet may require additional or different abilities than responding on the questionnaire itself.

2. Depending upon their prior experiences with them, respondents may find it more difficult to use a separate answer sheet than to respond on the questionnaire sheet.

3. It is normally more difficult and time-consuming for the respondent to use a separate answer sheet. However, separate answer sheets have been used successfully for some purposes.

4. When separate answer sheets are employed, the questionnaire booklets are reusable.

5. Respondents sometimes err in using the last spaces on a multiple choice answer sheet when there are more spaces than response alternatives. This can be avoided by the use of tailor-made sheets.

F. Use of Branching

Some questionnaires are constructed so that respondents need not answer every question in the survey. Branching is used to guide respondents through a survey instrument to appropriate questions. This technique requires the construction of questions which are integrated and then arranged to implement the purposes of the branching.

1. When to Use Branching

Branching is used when the researcher wants to screen respondents and assign them to subgroups. It is also a way to guard against having the respondents' answers be influenced by the question(s) that are bypassed in the branching (or branched around). This is known as position effect. The questionnaire educates the respondent on the topic which it covers. Branching can be used to measure the effects of the questionnaire educating the respondent. The more forward branching that occurs, the less education is being given to the respondents (they don't need it).

Branching can be used to reduce interview time for questionnaires administered by interview. Clear branching instructions are required for the interviewer, as well as interviewer training. Self-administered/group-administered questionnaires which use branching can reduce the time to complete for the respondents. However, there is a greater risk of item nonresponse following a branch for these questionnaires.

2. Filter Questions

Filter questions are developed to determine how respondents are to be guided through the questionnaire. Filter questions screen out respondents from certain sets of questions on the questionnaire. Branching is used so that the respondent can be routed into another subset of questions. Consequently, the survey functions as a set of filters through which some respondents pass while others are detained or routed into different topic area(s).

3. Branching Applications

Respondents are routed through the questionnaire by presenting them with more difficult or more concrete questions. This approach forces the respondent to consider the topic area from many viewpoints. Clear branching instructions are required for all questionnaires. Items immediately following a branch tend to have an increased rate of nonresponse; to this extent, the branching instructions were unclear or not understood.

4. Recommendations

Surveys conducted by interview which use branching may be choppy. Interviewers require smooth transitions between branches, and training in conducting the survey. Branching is best used to reduce interview time. Branching may be used to reduce/avoid exposing the respondent to items that are irrelevant or non-essential. This forces the interviewer to ask only pertinent questions regardless of the interviewer's persuasion. Branching for mail surveys and group-administered surveys has a greater probability of increasing item nonresponse rate than a survey conducted by interview.

There are alternatives to branching, such as the design of different questionnaire packages for the difficult categories of respondents. An illustration of this approach was used in the Army Research Institute's test of the Bradley Fighting Vehicle. Four separate questionnaires were designed: one for the driver, one for the track commander, one for the gunner, and one for the remaining personnel.

When branching questionnaires are used to measure respondent attitudes, there is a greater possibility for introducing bias into the data. Questionnaires covering topic areas dealing with fact instead of attitude are preferred when branching is used. Branching may be used to reduce/avoid exposing the respondent to items that are irrelevant or nonessential. This forces the interviewer to ask only pertinent questions regardless of the interviewer's persuasion. See Sections VI-F-1 and X-D-3.

Chapter X: Considerations Related to Questionnaire Administration

A. Overview

Considerations related to the administration of questionnaires are discussed in this chapter. Such matters are obviously of concern when questionnaires are constructed. Questionnaire instructions are discussed in Section X-B, anonymity for respondents in Section X-C, and motivational factors related to questionnaire administration in Section X-D. Administration time, characteristics of administrators, and administrative conditions are the topics of Section X-E, X-F, and X-G, respectively. The training of raters and other evaluators is the concern of Section X-H, while other factors related to questionnaire administration are considered in Section X-I.

B. Instructions

Care must be exercised in preparing instructions for questionnaires since they are quite likely to affect the way the respondent answers the questions. For example, even mildly anger-arousing printed instructions may elicit responses of negativism.

Although further research is needed to fully determine the influence of instructions on responses, some practical guidelines can be offered:

1. It is sometimes preferred that an oral statement of questionnaire purpose be given to respondents. If this is not practical or a person with appropriate credibility and/or status cannot be supplied to make the statement, then a printed statement must suffice. (See Section IX-D 3 regarding printed introductions.)

2. Lengthy instructions for completing questionnaires should be avoided. They may tend to confuse the respondents rather than help them.

3. The option of orally presenting instructions is often available. When oral instructions are given, they are usually given just prior to administering the questionnaire.

4. If instructions are given orally and an illustration is needed, a visual display should be available which may include a printed version of more complex instructions.

5. When questionnaires are group-administered, it should be announced that aides will check each respondent's questionnaire for completeness, if such a process can be implemented.

6. "Cute" examples on instructions should not be used. They will damage rapport and detract from the seriousness of the questionnaires, particularly for more mature and older respondents. It is best to use a neutral example that will be suitable for all respondents.

7. Obviously, instructions should be given in a way that all respondents can understand them. Care should be exercised about the level of vocabulary used.

An example is given on the following page of the instructions that might precede the items of a questionnaire. In this example, the responses were to be given on a separate "answer" or response sheet.

TRAINING ATTITUDE QUESTIONNAIRE (BASIC AND AIT)

INSTRUCTIONS: The purpose of this questionnaire is to obtain information from you regarding training, working and living while in the Army's Basic Training and Advanced Individual Training (AIT) program. Your answers will help the Army to determine what conditions are in need of improvement, and will assist the Army in determining the actions they must take to improve training and the quality of life for new soldiers in the Army. Your honest opinions are, therefore, essential.

We have no need to know who you are personally. No effort will be made to identify either you or your unit. DO NOT WRITE YOUR NAME, SOCIAL SECURITY NUMBER, OR UNIT on either the questionnaire or the answer sheet.

Each question should be answered by circling the letter on your answer sheet which is next to the answer which best describes your feelings. See sample question below:

SAMPLE QUESTION: 3. How old are you?

a. 17
b. 18
c. 19
d. 20
e. 21 or older

If you are 19 years old, you should circle the letter c on your answer sheet for question 3, as has been done below, since the letter c corresponds to your correct age of 19 on the questionnaire.

QUESTION NUMBER	RESPONSES (CIRCLE ONE)				
01	a	b	c	d	e
02	a	b	c	d	e
03	a	b	c	d	e
04	a	b	c	d	e

If you have any questions, please ask the questionnaire administrator for assistance. You will have 30 minutes to complete the questionnaire. You will all turn in your answer sheets, and leave at the same time. Do not turn the page and start to work until instructed to do so.

C. Anonymity for Respondents

1. Factors to be Considered

There are several factors to be considered when deciding whether to require the respondent's name or other identifying information on a questionnaire. Some of the factors are supported by research, while others are not.

a. If the respondents supplied their names, they are aware that they can be identified and called back. If respondents do not have to give their names or similar information, most will believe that they cannot be identified and called back for any type of accounting after their questionnaires have been collected.

b. The perception of anonymity seems to depend not only upon whether respondents give their names, but also on the conditions under which the questionnaires are administered. For example, paper-and-pencil questionnaires are more anonymous than structured interviews.

c. The effects of anonymity seem to be related to the content of the questionnaire. This is particularly true when information on sensitive areas is collected. For general attitudes, it may not matter.

d. The effects of anonymity may also depend upon who administers the questionnaire, and the circumstances under which it is administered. Responses may be distorted when respondents are identified and under high threat.

e. Respondents may be more lenient when rating other personnel if they think they will be identified.

2. Implications of the Privacy Act of 1974

If the experimenter, test officer, or questionnaire writer desires to obtain certain types of personal information from a respondent, the federal Privacy Act of 1974, in turn, requires that certain information first be given to the candidate respondent. One may use DA Form 4368-R, 1 May 75 for the purpose of communicating this information to the respondent. The form is shown filled out on page X-C 4. In this particular example, the research questions dealt with attitudes toward respondents' treatment in the Army.

A second example, Figure X-C-1, illustrates a more compact format. The same elements of information called for by DA Form 4368-R have been communicated; it's just that that form was not used.

A privacy act statement is not necessarily required as a part of all questionnaires that are administered to Army personnel. It is not necessary where only the personal information listed below is being requested. For example, no invasion of privacy is involved where soldiers are asked to evaluate some new/revised weapon, equipment, or organization regarding effectiveness and/or acceptability, and to answer any of the 12 items listed below. The collection or release of the following information does not require the consent of the respondent:

a. Grade.

b. Date of birth.

c. Date of rank.

d. Salary.

e. Present duty assignment.

f. Past duty assignments.

g. Future assignments (approved).

h. Unit and/or office address.

i. Unit and/or office phone number.

j. Source of commission.

k. Military and/or civilian education.

l. Promotion sequence.

Data collection procedures that guarantee anonymity are desirable for surveys. If the research methods cannot guarantee anonymity, then confidentiality of the data is to be protected. For operational test and evaluation research, participants should be informed that the data cannot be kept confidential. Surveys requiring the names of participants can have records coded as soon as possible. The key to the code can be stored for limited access to protect confidentiality. (See American Psychological Association (1982). Ethical principles in the conduct of research with human participants. Washington, DC.)

If you have any questions concerning the Privacy Act of 1974, you may obtain additional information from the ARI Field Unit. All questionnaire respondents must be advised of the requirements of the Privacy Act of 1974 when any of the 18 types of information listed below are being requested. This information can only be obtained from an individual on a voluntary basis. The release of any of the information listed below requires the prior and informed consent of the individual.

a. Name.

b. Social Security number.

c. Home address.

d. Home phone number.

e. Home of record.

f. Financial transactions.

g. Character quality.

h. Efficiency ratings.

i. Conduct ratings.

j. Legal affairs.

k. Religious preferences.

l. Number of allotments.

m. Amount of allotments.

n. Medical history.

o. Criminal history.

p. Fingerprints.

q. Voiceprints.

r. Photographs.

DATA REQUIRED BY THE PRIVACY ACT OF 1974
(5 U.S.C. 552a)

TITLE OF FORM	PRESCRIBING DIRECTIVE
	AR 70-1

1. AUTHORITY

10 USC Sec 4503

2. PRINCIPAL PURPOSE(S)

The data collected with the attached form are to be used to research purposes only.

3. ROUTINE USES

This is an experimental personnel data collection form developed by the U.S. Army Research Institute for the Behavioral and Social Sciences pursuant to its research mission as prescribed in AR 70-1. When identifier (name or Social Security Number) are requested they are to be used for administrative and statistical control purposes only. Full confidentiality of the responses will be maintained in the processing of these data.

4. MANDATORY OR VOLUNTARY DISCLOSURE AND EFFECT ON INDIVIDUAL NOT PROVIDING INFORMATION

Your participation in this research is strictly voluntary. Individuals are encouraged to provide complete and accurate information in the interests of the research, but there will be no effect on individuals for not providing all or any part of the information. This notice may be detached from the rest of the form and retained by the individual if so desired.

FORM Privacy Act Statement - 26 Sep 75

DA Form 4368-R, 1 May 75

Figure X-C-1

An Example of a Privacy Act Statement

11B/C GRADUATE FIELD SURVEY
(Prescribing Directive: AR 600-46; TRADOC Ltr dtd 29 Aug 75)

INFORMATION PRIVACY ACT STATEMENT

1. Authority: 5 USC 301, 10 USC 3012, Authority for the Secretary of the Army to Issue AR's; 44 USC 3101, Authority for Collecting Necessary Data.

2. Principal Purpose: To collect data to evaluate the effectiveness of individual training received prior to joining one's initial unit of assignment.

3. Routine Uses: The data collected with this form are to be used for research purposes only. They will not become a part of any individual's record and will not be used in whole or in part in making any determination about an individual.

   The identifiers (name or Social Security Number) are to be used for administrative and statistical control purposes only. Full confidentiality of responses will be maintained in the processing of these data.

4. Mandatory or Voluntary Disclosure and Effect on Individual Not Providing Information: Voluntary - Your participation in this research is strictly voluntary. Individuals are encouraged to provide complete and accurate information in the interests of the research, but there will be no effect on individuals not providing all or any part of the infomation.

   This notice may be detached from the rest of this form and retained by the individual answering the questionnaire if so desired.

D. Motivational Factors

This section considers the effects of lack of motivation, and some ways of providing a desirable level of motivation to respondents during the questionnaire administration process.

1. Effects of Lack of Motivation

Generally, the results of any study will suffer distortion if those to whom the questionnaire is distributed are not sufficiently motivated. If they have the choice, they will not respond at all. If they do have to respond or are just minimally motivated, they may omit items, make patterned or random responses, or just generally respond poorly. As a result, the reliability and validity of the responses will be decreased, and the results of the study would lead their reader/user into some degree of error.

2. Ego Involving Potential Respondents in the Study

There are a number of ways that motivation can be increased by ego involving potential respondents. Some of the ways are given below:

a. The special role of the respondent in the study can be emphasized.

b. Responsibility can be stressed when it is appropriate to do so.

c. The wording of cover letters, if used, affects ego involvement. Help may sometimes be requested on the basis of appealing to the self-interests of the respondent. There is evidence that this type of appeal helps most with less educated respondents.

3. Stimulating the Return of Remotely Administered Questionnaires

Obviously, whatever involves the egos of potential respondents in a study also stimulates the return of remotely administered questionnaires, such as those distributed by mail. Other ways of stimulating the return or response rate are:

a. Return rates may often be significantly improved when a letter is sent in advance notifying the potential respondents that they will receive a questionnaire and their help is needed in filling it out.

b. Stamped and return addressed envelopes can be sent with the questionnaire. There is evidence that this does increase response rate.

c. There is contradictory evidence about whether short questionnaires are returned more frequently than longer ones, but one would probably believe this to be true.

d. Follow-up reminders can be sent to those who do not promptly return their questionnaires. There is some question, however, regarding how much such follow-ups increase response rate. At times, it may not be cost effective, so maybe the decision should be a function of whether or not the initial return rate was adequate.

e. Telephone interviews and face-to-face interviews generally have a higher response rate than mail surveys.

f. Response rate for telephone interviews can be increased by changing the format from what would be used in a face-to-face interview. Select fewer items and items which are shorter in length for telephone interviews to reduce telephone disconnects.

g. Nonresponse for items following a branch may increase the overall item nonresponse rate, especially for mail surveys. If branching can be avoided, this may increase item response rate.

4. Use of Incentives

The evidence has been mixed regarding the extent to which motivation is increased through the use of incentives. Incentives may include money, time off, special privileges, etc. Generally, however, it is agreed that incentives usually help increase the response rate with remotely administered questionnaires.

5. Other Motivational Factors Related to Questionnaire Administration

Many additional motivational factors related to questionnaire administration can be noted or inferred from other sections in this manual. Some of them are:

a. Respondents often have preferences for certain item formats, although sometimes such preferences may not offer any advantage in terms of reliability and validity. Some subjects prefer rating scales to forced choice items. With forced choice, some like the option of indicating the degree of applicability of each statement. Some do not like forced choice Q-sort (see Section IV-H). Some prefer multiple category to two category options. In some studies, Likert scales have been preferred to behavioral scales. Behaviorally Anchored Rating Scales have been preferred to Mixed Standard Scales, etc. These preferences may relate to familiarity of the respondent with given item types. There is not much that the questionnaire designer can do about such preferences, except to note that they exist.

b. Researchers in recent years have explored the cognitive complexity of respondents to match them to formats which are cognitively compatible. There have been problems with replication for this research.

c. Motivation may be increased by offering feedback of study results to the respondent.

d. Every effort should be made to praise the respondents or potential respondents, to the extent that it is reasonable.

e. Long, vague, or boring questionnaire sessions should be avoided, since it will decrease respondent motivation to continue attending and providing "best" responses.

f. Questionnaire administration sessions should not be scheduled when there are conflicts with other activities of greater interest to the respondents. Nor, in general, should they be scheduled very early or very late in the day.

g. Volunteers are usually more motivated to fill out questionnaires than are nonvolunteers. However, their replies may be more biased.

h. When respondents are told that they may leave as soon as they have completed the questionnaire, they usually do a much more hasty and unsatisfactory job than when they are given a specific time for completion, and are told that they cannot leave until the time period is up.

i. See Chapter XIV about the behavior of interviewers.

E. Administration Time

Little is known about the effects of questionnaire administration time on respondents' motivation, or of the effects of setting time limits for completing questionnaires. The questionnaire administration period should generally have been determined in advance by pretesting. There will be some variability in the length of time taken to complete a questionnaire. There is remarkable consistency among those who are sincere in attempting to do an accurate and complete job of answering all questions.

When a questionnaire is administered to a group of respondents, the instruction should emphasize that all respondents will be given plenty of time to answer the questions. As indicated earlier in X-D 5 h, the instructions should not tell the respondents that they can leave as soon as they have finished the questionnaire. Many will then cut short their efforts to answer the questions. There is little hope of obtaining carefully considered evaluative responses on a questionnaire if the respondents knows that the faster they finish the questionnaire, the sooner they will be able to go home.

Questionnaire administration time is obviously related to questionnaire length, which is the topic of Section IX-C.

One should try to determine empirically the maximum time needed to complete a given questionnaire. If the questionnaire is group-administered, the maximum time for the slowest respondents should usually be used in scheduling the administration of the questionnaire.

F. Characteristics of Administration

Little has been established in the research literature about how the characteristics of questionnaire administrators affect the overall process with nonremotely ' inistered questionnaires. The following items may be noted:

1. In most cases, it is felt that the sex of the administrator has no effect on the responses received. There may, however, be certain motivational effects.

2. The military rank of the administrator may have an effect on the respondent, but no research has been performed to examine this.

3. Any effect that the race of the administrator has on the respondent may also be a function of the content material of the questionnaire, e.g., race would be expected to influence responses on a race relations questionnaire more than on a questionnaire dealing with rifle comparisons. The effects should probably be viewed as the result of interaction between administrator and respondent characteristics, and the questions being asked.

4. Implications exist for biasing survey results whenever surveys incorporate face-to-face interviewing with individuals from different ethnic backgrounds. Items with racial content used in a questionnaire are especially sensitive to such biasing.

5. See Chapter XIV about the influence of an interviewer on the interviewee.

G. Administration Conditions

Questionnaire administration conditions obviously cannot be controlled with remotely administered questionnaires. With group-administered questionnaires, the following guidance is offered:

1. Administration conditions should be provided which are most appropriate to the particular type of respondent completing the questionnaire.

2. Administration conditions have an effect on questionnaire responses. For example, different responses may be obtained if the questionnaire is filled out in a group situation on the job rather than individually at home.

3. When personnel are being rated, different ratings may be obtained, depending on how acquainted the rater and ratee are.

4. For Army field test evaluations, the circumstances under which questionnaires must/can be administered will vary rather widely. There may be times when no writing surface(s) or pencils are available; clipboards and pencils should be supplied if this problem can be anticipated. If the needed materials cannot be brought to the respondents, then arrange to move them to a place where the materials and other environmental conditions are satisfactory.

5. Respondents should be required to give their answers without being influenced by other respondents. Achieving this requires respondents to be somewhat separated and/or to have the administrator(s) watching them. Simply instructing them not to consult with each other is usually not sufficient.

H. Training of Field Test Evaluators

An extended discussion of the training of raters and other test evaluators is not undertaken in of this manual. The following suggestions, however, can be offered about the general training of the Army field test evaluators. See Section X-B regarding questionnaire administration instructions.

1. Impress on test evaluators that they are supposed to answer the questionnaire based upon what they observe in the test. Stress the need for evaluations based only upon what was seen during the test exercise, regardless of any personal feelings or knowledge of concepts or equipment as might exist in a true combat environment (except in special instances where this is specifically asked for). To help identify and reduce prejudgment, a broad question might be included to permit the evaluators to express any biases they may have. It may be a question such as "Based on your personal experience, do you feel the "DPST" is a useful approach to real daily problems, i.e., outside a test exercise environment?" Such a question would permit the evaluators an outlet for preconceived opinions and attitudes which otherwise would color their view of the events observed during the exercise. On the other hand, in some situations the evaluators might feel it necessary to defend their personal judgment by biasing their answers to the remaining question answers!

2. Stress the importance of evaluators to the success of the test. Perhaps briefly indicate some actions which have been taken to implement concepts supported by evaluative data from previous tests.

3. Permit evaluators (particularly after the pilot test) to sound off about the forms and their perceived inadequacies, regardless of how unreasonable these complaints might be. The goal is to have all evaluators answering questionnaires understand that they are active and important contributors rather than just a means of satisfying some obscure test requirement.

4. Examine all turned-in questionnaires to ensure that they have been filled out and understood. This procedure should continue throughout the entire series of tests.

5. Stress the notion that complete honesty and objectivity is needed. Sometimes evaluators try to please the test sponsors, to the detriment of the test.

6. Indicate to evaluators, perhaps on the top of all questionnaires or verbally, that they may make marginal note clarifications concerning their scale value selection for any rating question. This will increase posttest accuracy in determining questions which are scaled awkwardly or unclearly stated. This is particularly crucial during the pretesting or pilot test. Notes should be made regarding question structure immediately as they occur to the evaluator or the difficulty is likely to be forgotten.

7. Prior to having the evaluators complete questionnaires, ask all or a sample of randomly selected evaluators to orally describe to the other evaluators what they believe each question is asking. This procedure will reduce differences between judges because of varying semantic interpretations. By the time of the actual exercise, all evaluators should generally agree, for example, on the meaning of "command and control effectiveness," "fire power potential," etc. If this is done, the criteria will have mutual acceptance. This procedure is also useful during the pretest to assist in the selection of item wording that will be understood by the respondents.

8. Evaluators should be forewarned about biases such as the halo effect, central tendency, and others discussed in Chapter XII. If it is explained to the evaluator that these are common biases to which we are all subject, the evaluators will be better able to consider the fairness and accuracy of their observations. Training to reduce rating errors is especially effective when the training is extensive, and allows evaluators to practice. Evaluator experience with the questionnaire may improve rating accuracy. Short training programs may have little impact on rating quality. To train evaluators, effective training should include observational techniques in conjunction with written performance observations between rating periods.

9. The independent, non-collaborative, evaluation of each question should be stressed.

I. Other Factors Related to Questionnaire Administration

Some other factors related to questionnaire administration that have not been discussed in other sections of this manual are addressed below:

1. Respondents may at times be influenced by the title of the questionnaire. The word "test" should not be used in a title of a questionnaire as it may imply that it is a test of the respondent's knowledge.

2. A problem with Army field test evaluations concerns undue influence by the questionnaire administrator. It is sometimes necessary to use line officers from the units of the test subjects as questionnaire administrators. When outside administrators are used, they must be carefully instructed to make no comments whatsoever regarding their personal opinions of the items being evaluated. An offhand comment by a company commander administrator to his/her company regarding the "goodness" or "badness" of a piece of equipment or concept being evaluated can exert an influence sufficient to distort the results significantly from what they would otherwise have been.

3. The manner in which test subjects are selected and utilized in operational tests may affect the manner in which they respond to questionnaire items. For example, separate groups with no prior experience with either the test system or the current standard system could evaluate each system. This would exclude pretest biases, but test subjects would have no basis to compare the two systems. Alternatively, the same group of test subjects could use both systems in rotation. However, this procedure may result in a bias for or against one or both systems as a function of which was used first. In this respect too, personnel having extensive prior experience with a current standard system may introduce their pretest biases for or against that system when it is being evaluated against a candidate replacement system. The consequence of such considerations is that the type of system evaluation intended will govern the way evaluators and/or test subjects are selected and utilized. The methods of selection and utilization will influence the way questionnaires must be designed, and in turn suggest the types of problems likely to arise.

Chapter XI: Pretesting of Questionnaires

A. Overview

Even the most careful screening of a questionnaire by its developer or by questionnaire construction experts will usually not reveal all of its faults. Pretesting is an important and essential procedure to follow before administering any questionnaire. Its purpose is, of course, to find those overlooked problems and faults that would otherwise reduce the validity of the information obtained from the questionnaire responses. However, just any pretest will not do. One must know how to pretest the items and what to look for.

Some guidelines for pretesting questionnaires are given in this chapter. Pretesting may seem to some uninformed individuals to be a waste of time, especially when the author may have asked several people in his/her own office to critique the questions, or perhaps even asked a questionnaire specialist to critique it. However, pretesting is an investment that is well worthwhile. It is crucial if the decision that will result from the questionnaire is of any importance.

B. Guidelines for Pretesting Questionnaires

1. Before a pretest is conducted and a questionnaire is constructed, hypotheses and questionnaire items are developed. The hypotheses are presented to a group of individuals who are subject matter experts. The group performs a preliminary assessment of the hypotheses and items. Modification may be required regarding the hypotheses and/or questionnaire items.

2. Initially, open-ended questions are established and placed into a logical sequence. Pretesting may provide information that can be used to convert open-ended questions to multiple choice questions to facilitate data reduction and analysis. Instructions are developed to accompany the questionnaire, and they are included as part of the pretest. If branching is used, it should be kept to a minimum.

3. It is important that the respondents employed in pretesting be representative of the eventual target respondents. For example, if infantry enlisted men will perform in a test and then take the questionnaire, it should not be pretested with respondents who are armored officers; even infantry officers would not be satisfactory.

4. The pretest is more useful if it is conducted by someone who knows the operations to be performed in the test and who also knows the subject matter that the questionnaire covers. It is best if the question writer is knowledgeable about these operations and conducts the pretest.

5. Early versions of questionnaires may contain instructions, item stems, response alternatives, and item ordering that are confusing to respondents. It is possible that more than one pretest will need to be conducted. Some researchers have been known to conduct up to six or more pretests.

6. Interview and pretest some of the pretest respondents one at a time or in a group. Ask each respondent to read each question and explain its meaning. Also ask them to explain the meaning of the response alternatives, and to make their choice. Ask the respondent to explain why a particular choice was made. The respondents' answers will frequently reveal incorrect assumptions and possible rationales that the question writer never dreamed possible. They will also help to identify lack of understanding of particular words, vague or ambiguous phrases, ill defined or loaded questions, inadequate space for recording answers, inappropriate sequencing of items, etc.

7. One good technique for pretesting is to have respondents complete the questionnaire. A discussion can then be had where respondents read each question aloud and then tell you what it means. Any difficulties at all should be a cause for concern and revision. Pretest methodology can be strengthened if discussions are tape recorded, and suggestions for modifications are systematically coded. This is especially useful when pretesting a questionnaire that will be administered by interview.

8. During pretesting, the respondents should be encouraged to make marginal notes on the questionnaire regarding sentence structure, unclear questions or statements, etc. Pretests will provide a good idea as to the length of time it takes to complete the questionnaire.

9. When attitude questions, especially, are being pretested, individuals who may hold minority views should be included. This will help identify loaded questions.

10. Pretests for the selection of verbal anchors are valuable in building rating scale content validity and reliability. Rather than employing anchors which seem appropriate, the anchors used in the final scales should be selected as a result of analyses of pretests of respondents similar to those who will be participating in the final test.

11. While pretesting a questionnaire, a high proportion of respondents giving no response or a "Don't know" response should be a cause for concern. However, a low number of "Don't know" responses (especially for multiple choice items) does not guarantee that the question is good.

12. After pretesting, each question should be reviewed and its inclusion in the questionnaire justified. Questions that do not add significant information or that largely duplicate other questions can profitably be eliminated. Quantitative item reduction techniques will depend on the type of scale that is being used, e.g., Thurstone, Likert, Guttman, etc. A discussion of quantitative item reduction techniques is outside the current scope of this manual. Army personnel may check with the Army Research Institute Field Unit closest to them for help in this area.

Chapter XII: Characteristics of Respondents That Influence Questionnaire Results

A. Overview

This chapter discusses some characteristics of respondents that influence questionnaire results. It therefore identifies some of the principal sources of error in the reporting of observations and/or the evaluation of performance in, for example, operational Army field tests. Additional research is required, however, to determine their relative contributions to error variance.

Sections XII-B and C present a discussion of various biases, response sets, or other sources of error. There is some confusion in the literature regarding the use of these terms, but they are similar. A bias is: a tendency to deviate from a true value; a tendency to favor a certain position or conclusion; or an attitude either for or against a certain unproved hypothesis which prevents an individual from evaluating the evidence correctly. A response set or response bias refers to the tendency of a respondent to answer questions in a particular way almost independent of the content of the questions. An error is simply a mistake or departure from correctness.

Section XII-D addresses the effects of attitudes of respondents on questionnaire results, while Section XII-E considers the effects of demographic characteristics on responses.

One of the main purposes of this chapter is to alert the questionnaire designer to some of the characteristics of respondents that influence questionnaire results. There are ways that some of the biases and errors can be controlled, but not all of them. And there appears to be no easy way of detecting the influence of a response set nor of neutralizing it. More detailed identification and control methods are areas of needed further research.

B. Social Desirability and Acquiescence Response Sets

Social desirability is a response set where persons answer according to the norms they believe society supports. It is the tendency to agree with items the respondents believe reflect socially desirable attitudes in order to show themselves in a better light. Acquiescence response set is the tendency to consistently agree, to say "Yes," or to say "True." It is a general tendency to assent rather than dissent. Although there have been a number of studies about each, a detailed discussion of them is beyond the scope of this manual. (See P-77-2, Questionnaire Contruction Manual Annex, Literature Survey and Bibliography; and P-85-j, Questionnaires: Literature Survey and Bibliography.) Some comments about each are presented below.

1. Social Desirability Response Set

a. Social desirability response set seems to operate whenever the respondent has the opportunity to respond in terms of it. Some believe that its effect is so powerful that respondents would not tend to deviate from social norms in their answers even though their behavior denied what they said.

b. Several authors have identified respondents with a high social desirability response rate. They found these respondents to give more true responses to neutral items, to be more susceptible to social pressures, to more likely be introverts, and to score higher on a "lie" scale.

c. Faking or responding with socially desirable answers which are not true is part of the response set.

d. Anonymity fails to eliminate the social desirability response set.

e. The forced choice instrument format has been studied for its susceptibility to social desirability response set, a factor it was intended to control. Some authors found the forced choice method minimized the effects of social desirability, while others think the factor still needs additional control. One study concludes that in comparing different forced choice formats, ambiguous items tend to be freer of social desirability response set than positively or negatively worded items. In any case, the evidence indicates that the social desirability problem is sometimes less in forced choice formats than in other item types such as graphic rating scales. Forced choice formats may or may not reduce bias.

f. Card sorts also need control to eliminate social desirability bias.

g. Respondents may be confounding trait dimensions with response alternatives on clinical instruments. There is some evidence that respondents have a stronger tendency to select response alternatives opposite in desirability when a socially undesirable response alternative is presented first.

h. Procedures have been developed for controlling or balancing social desirability by using loaded items in the questionnaire and then adjusting the respondent's score. The social desirability score from the loaded items can also be correlated with each of the other items on the questionnaire. The responses on those items with a statistically significant correlation can then be corrected by moving the response one or more steps from the socially desirable response to give a more accurate result.

2. Acquiescence Response Set

a. The acquiescence response set is defined as a behavioral attitude by the respondents to agree and accept, even if they must alter their original opinions to do so.

b. The acquiescence response set seems to operate especially when statements are in the form of plausible generalities.

c. The response set may occur more with difficult than with easy questionnaire material.

d. Acquiescence response set may be a personality trait.

e. There is a concern that social desriability and acquiescence response sets may be related in such a way that an individual with a tendency toward conformity will consistently reflect both biases.

f. Controls for acquiescence response set have been researched. Stating the question stem in a neutral manner may help minimize acquiescence. The effects of acquiescence response set may also be partially controlled by using two alternate questionnaire forms with the question stated positively on half of the forms and stated negatively on the other half. The balancing of scales (e.g., equal number of positive and negative points) may also be of value in counteracting acquiescence.

C. Other Response Sets or Errors

This section notes a number of other response sets or errors of which the questionnaire developer should be aware.

1. Error of Central Tendency

Some respondents tend to avoid endpoints on a scale, and pick a middle value regardless of their true feelings. It may be more common when the respondents are not very familiar with whatever they are being asked to rate. It may be counteracted by adjusting the strength of the response alternatives so that there are greater differences in meaning between alternatives near the ends of the scale than between alternatives near the center.

a. In one study, responses tended to be toward the center of the scale when item length increased (more than 17 words). Respondents selected response alternatives toward the positive end of the scale when item length was short (less than 17 words). Items may be ambiguous to the respondent when they are long and negatively worded. This appeared to influence respondents to rate items toward the mid-range of the scale.

2. Extreme Response Set

On the other hand, some individuals tend to consistently select exaggerated choices for positions. It can be recognized when a respondent makes a pattern of answers which tend to be unevenly distributed toward one or both ends of a scale. Research indicates that this response set may be a personality characteristic.

a. Research evidence indicates that positively worded items receive higher mean responses than negatively worded items. There is the possibility that respondents prefer or agree with positively worded items, and rate them higher.

b. For cross-cultural survey research, there is some evidence that response style may vary from country to country. One study concluded that there was a tendency by respondents in the Philippines to use a positive response style, and by respondents in Italy to use a negative response style.

3. Halo Effect

Halo effect was originally defined as a tendency when one is estimating or rating a person with respect to a given trait, to be influenced by some other trait or by one's general impression of the person. It is, however, also applicable to ratings of other than people. For example, if field test evaluators know that a particular weapon system did well in one phase of a test, they may be influenced to give high ratings to the system in later test phases - even when the system performs poorly.

a. Most studies of ways to control halo effect have dealt with ratings of traits of personnel by other personnel, a matter not of great concern in this manual. The forced choice technique and Mixed Standard Scales minimize halo effect in some situations. Ratings will also be less distorted if questionnaire items are constructed so as to relate to clearly observable aspects of behavior which do not overlap. It is doubtful that the influence of halo effects can be completely eliminated from the responses to any questionnaire.

b. Behavioral scales such as Behaviorally Anchored Rating Scales (BARS), Behavioral Expectation Scales (BES), and Mixed Standard Scales (MSS) have been developed to measure performance. There is evidence that the use of behavioral scales in conjunction with intensive training can reduce halo error. This combination of behavioral scale and training appears to be more effective than graphic rating scales, trait scales, and Likert scales in reducing halo error. The length of the training session appears to influence whether halo error will be reduced. Training sessions of 5-minute duration have had little impact on the quality of ratings. Training sessions of 3-hour duration were found to reduce halo error. Intensive training sessions may not reduce other types of rating errors even though they tend to reduce halo error.

4. Leniency Error

Leniency error refers to a general, constant tendency for a rater to rate either too high or too low in the direction of being too generous. It appears similar to halo effect, except that it is independent of the trait or factor being rated. Some raters have an opposite tendency to rate too severely. In one study, respondents rated Likert scales with less leniency error than they rated behavioral scales. These findings may not be consistent across studies. In large groups of raters, the opposite tendencies should balance out.

5. Logical Error

Logical error is also similar to halo effect. It is due to the fact that raters are likely to give similar ratings to traits or items that seem logically related. For example, field test evaluators may know that a counterattack was extremely successful; they may therefore reason that command and control was also very effective and should receive an equivalent high evaluation because a successful counterattack is a function of good command and control. Such reasoning assumes a dependence which may or may not be true. Logical error may be avoided in part by asking for judgments of objectively observable actions or behavior.

6. Proximity Error

Proximity error occurs when, due to the ordering of questionnaire items, the answer to one item results in an answer to a subsequent question being substantially changed from what it would otherwise have been. Little is known about its influence in field test situations; most research in this area has concerned the rating of personality trait variables.

7. Contrast Error

Contrast error refers to a tendency of raters to rate others in the opposite direction from themselves in regard to a trait. Little research has been done on this source of error.

8. Feedback Bias

Research shows that if observers are informed of experimental hypotheses, and if they receive daily feedback indicating how well their data support the hypotheses, they will tend to report data supporting those hypotheses - even when the reverse is true! This bias does not seem to occur, however, when observers are informed only of the experimental hypotheses with no follow-up. Taking precautions to assure high levels of observer accuracy minimizes the bias.

D. Effects of General Attitudes of Respondents

Limited research has been conducted upon how the attitudes of a respondent influence questionnaire results. The following, however, should be noted:

1. Respondents at times base their ratings not on what is observed but on what they believed prior to the observation. Beliefs and opinions may affect results.

2. It is generally believed that judges used as part of the process of determining scale values can rate items without being influenced by their own attitudes. There is also some evidence to the contrary.

3. Unstable or changing responses to questionnaires may be caused by shifts in the mood of the respondent, relative values among the possible choices, and the degree of interest present in the question.

4. As questions become more ambiguous, responses normally become more influenced by attitudes.

5. It may be desirable to revise a questionnaire when norms of groups differ greatly from those with whom the questionnaire was pretested or previously administered.

E. Effects of Demographic Characteristics on Responses

Demographic characteristics have been shown to influence questionnaire results. Similarities of such variables among respondents often tend to be related to a response pattern. These variables include: age, religion, sex, intelligence, marital status, parenthood, socioeconomic class, nationality, race, urban or rural residence, income, rank and experience. Questionnaires should, therefore, be designed with the respondents background in mind. When there is a suspicion that demographic characteristics may affect response, the data should be analyzed by type of respondent.

1. Research indicates that the racial background of survey interviewers does not seem to affect survey results when the questions do not deal with racial stereotypes and are not threatening. For most questionnaires administered by interview, it would be possible to assign interviewers of different racial backgrounds regardless of respondents' racial backgrounds.

2. Racial background has been known to influence rating errors on performance measures. However, this phenomenon has not been observed consistently.

3. Survey items which tend to be most sensitive to differences in response pattern by gender are those dealing with sex role stereotypes. Items that are relevant to technical background experience which females may not have, may yield gender-related response patterns.

4. It was hypothesized in one study that females would rate items according to social desirability. It was suggested that females have a greater need for social approval, that they are more impulsive than males, and that this would be reflected by male/female differences in rating. The results indicated that there were no significant differences between ratings by gender. Some studies have found differences in rating by females, while other studies have not. The overall effect of gender differences in response pattern has little support. Rating characteristics identified by gender are usually not enough to explain rating differences. Other variables must be taken into account as well, such as education, race, age, etc.

5. Respondents with low levels of education may be confused by items constructed in absolute terms. This may result in rating the items with an inconsistent response pattern. One should review questionnaire items to ensure that the content is not ambiguous to respondents with a low level of education.

6. Survey items which request an opinion regarding an obscure topic area may elicit a "Don't know" response by respondents with higher levels of education. Respondents with higher levels of education seem to be more willing to admit they do not have knowledge of obscure topics. For this type of question, respondents with less education have a tendency to give an opinion. Respondents with low levels of education do not appear to admit they don't know, but instead select a response alternative to represent their opinion.

7. For questions which are not obscure, the "Don't know" response alternative may be selected most frequently by respondents who have the least amount of education. Individuals with less education appear to be the most influenced when a "Don't know" response alternative is included.

8. The age of respondents does not appear to influence their ability to use different types of rating scales. However, the educational level of respondents may affect the way in which different scales are rated.

9. The content of some items may be related to the historical perspective of different age groups. For such items, the responses may be associated with different response patterns according to the age of respondents.

10. Nonresponse to an entire survey or to specific items in a survey remains a threat to the validity of survey results. Research indicates that nonresponse rates are sometimes associated with age of respondents. Item nonresponse rate may be reduced by eliminating branching from surveys that include respondents approximately 60 years and older.

11. When surveys are conducted by interview, it is important to be sure that older interviewers (about 50 or over) are following the standard format and interview guide. Of course, all interviewers need to conduct standardized interviews. One study indicated that older interviewers made more errors by conducting the interview in a non-standardized way. Possibly, the interviewers felt that their years of experience afforded them the opportunity to probe questions more thoroughly, and to somewhat modify the interview guide as they progressed through the interview.

Chapter XIII: Evaluating Questionnaire Results

A. Overview

An extended discussion on evaluating questionnaire results is currently outside the scope of this manual on questionnaire development. However, Army personnel may check with the Army Research Institute-Field Unit closest to them for help in the areas of coding and data analyses. There are some factors relating to the evaluation of questionnaire results that should be noted since they may influence how questionnaires are designed and developed. Section XIII-B considers the scoring and coding of questionnaire responses, and Section XIII-C contains some notes about data analyses.

B. Scoring Questionnaire Responses

1. Practical Considerations

a. Both time and money can be saved by planning the questionnaire in line with scoring and tabulation requirements. The phrasing of questions and their sequencing and layout affect tabulation time. For example, it is advantageous to have data coded and entered for analysis directly from edited questionnaires. Questionnaires consisting of only closed-end items will have a lower level of error for data entry than open-ended items. This is a more cost-effective approach. However, there are some drawbacks such as greater difficulty in verifying the coding and greater data entry time than when using a coding sheet.

b. A decision should be made ahead of time regarding whether the data will be tabulated by hand or machine.

c. Response alternatives should be precoded whenever possible. Codes for open-ended items are more difficult to construct than codes for closed-end items. To develop open-ended item codes, list out possible responses to the item. Pretest the questionnaire to classify responses to open-ended items. Construct a classification system and code. Pretest the code and revise as necessary. Develop a separate code for responses that were not possible to fit into the classification system above.

d. Codes need to be developed which guide coders in assigning code numbers to each answer. This includes the following: codes for missing data for item nonresponse, codes for item responses that are uncodable due to poor respondent performance, and a code for the "Don't know" response alternative.

e. Code books are constructed to define, clarify, and amend codes used during the coding process. Codes that have caused difficulty for the coders should be noted, such as classification systems and codes for open-ended items. Coders require training on specifics of the classification system and codes used for the study, and for the general principles of coding.

f. Since it does not seem to matter if items are scrambled or in blocks according to content, blocking may be preferred due to greater hand scoring ease.

g. Telephone surveys now use Computer Assisted Telephone Interviewing (CATI). These systems are still in experimental stages, and they require extensive programming. Items are read off the CRT screen, and telephone interviewers type respondent answers into a terminal for direct data entry.

h. See Section IX-E regarding the use of answer sheets.

2. Other Considerations

a. There may be a justification for scoring rating scale items dichotomously according to the direction of response. It is sometimes done when bipolar scales are analyzed in terms of the proportion of responses in either direction of the basic dichotomy. The justification is based upon results that seem to indicate that composite scores reflect primarily the direction of responses and only to a minor extent their intensities.

b. One investigator found that many Likert-type rating scales consisting of 2 through 19 steps may be collapsed into two or three measurement categories for analysis with no lack of precision.

c. When working with paired comparison items with a "No preference" option, the "No preference" responses can often be either divided proportionate to the preference responses, or disregarded altogether. The basis for this suggestion is that respondents who claim neutrality appear to exhibit the same preference patterns as those who express a preference.

d. By using any one of several methods of scoring or transforming self-rating scale raw scores, it is usually possible to approximate dichotomous forced choice results with considerable saving in administration time, and a small gain in test-retest reliability.

e. Investigators sometimes use intensity scores as well as rating scale content scores. One way of obtaining an intensity score is to follow each question with the query, "How strongly do you feel about this?" A second way involves weighting extreme responses (positive and negative) as 2, moderate responses as 1, and neutral responses as 0. These weights can then be summed for an intensity score.

C. Data Analyses

A detailed discussion of data analyses is beyond the scope of this manual; however, some basic data analysis issues have been mentioned in related chapters. Additionally, the following points are also noted:

1. Analyses of questionnaire responses are chiefly of two types: summary tabulations and statistical analyses. Tabulations are used primarily for the presentation of results. Statistical tests are used to determine whether the differences in the results are significant. Statistical literature is available which presents numerous tests usable in such analyses.

2. As part of the questionnaire development process, tentative (dummy) analysis tables should be developed to assure that the data to be obtained are appropriate.

3. Weights can be assigned to questionnaires when there is a probability that the selection of respondents is not representative of the population as a whole. For example, a sample distribution drawn from a list of service personnel receiving training, and enrolled in various courses, may result in unequal probability sampling. Since the subjects may be enrolled in more than one course, the more courses they take, the greater the chance they will be selected into the sample.

   Weights are also used in making adjustments for total nonresponse and in poststratification. They are able to assign greater importance to some sampled elements than to others in the data analysis. Poststratification conforms the sample distribution to the known population distribution. The sample distribution is adjusted across the strata. This is useful when the population is known, but the stratified sample elements cannot be determined at the selection stage. In such situations, prior stratification is not employable, although poststratification may be applied later. When a sample is weighted to a known population, it will adjust for the sampling fluctuations, as well as for nonresponse. For example, if nonresponse is higher for a specific age group, the sample will conform to the known age distribution when weighted. The development of weights is a difficult task. Standard computer programs for weighted data can be applied in data analysis.

4. Four kinds of measurement scales have been identified: nominal, ordinal, interval, and ratio. Appropriate statistical analyses are associated with each. Hence, the data analysis limitations of various forms of questionnaires should be considered before an instrument is designed. For example, less can be done statistically with open-ended questions than with ranking questions.

Chapter XIV: Interview Considerations

A. Overview

If properly used, the interview is an effective means of obtaining data. It is a technique in which an individual is questioned by a skilled and trained interviewer who records all replies, preferably verbatim in most cases. Most of the principals of questionnaire construction discussed in previous chapters pertain to the interview as well. This chapter, however, notes some issues specifically related to interviews.

Section XIV-B presents the distinction between structured and unstructured interviews. Interviewer's characteristics relative to the interviewee are noted in Section XIV-C. Situational factors are noted in Section XIV-D, while the topics of Sections XIV-E, F, and G are, respectively, training interviewers, data recording and reduction, and special problems. There is, unfortunately, little that can be recommended to avoid some of the problems noted in this chapter. The questionnaire developer should, in any case, be aware of them.

B. Structured and Unstructured Interviews

The term "structured" when applied to interviews is intended to emphasize that the interviewer employs a script of all the questions to be asked. In the unstructured interview, the interviewers may know many of the topics to be covered but they need to learn more about the subject overall, so they are willing to be led by the interviewee even into digressions. Unstructured interviews may occur as a preliminary to preparing either a questionnaire or a structured interview script. One could use a questionnaire as the script for a structured interview if one already had the questionnaire developed, but not enough time to convert it to a more convenient format. The main difference between the structured interview and questionnaire is procedural.

The degree of proficiency required of interviewers in conducting an unstructured interview is generally not available during Army field test evaluations. A structured interview requires the interviewer to have only moderate skill and proficiency, and hence is usually preferred. The advantages of the structured interview include: the opportunity to probe for all the facts when the respondent gives only a partial or incomplete response; a chance to ensure that the question is thoroughly understood by the respondent; and an opportunity to pursue other problem areas which may arise during an interview. The structured interview is almost always preferable to a questionnaire when the test group is small (10 to 20), and when time and test conditions permit.

As noted in Section II-B, unstructured interviews are not included within the definition of questionnaire used in this manual. They are, therefore, not discussed further.

C. Interviewer's Characteristics Relative to Interviewee

More research is needed to identify how characteristics of an interviewer affect the respondent. Some areas of concern are presented below.

1. Rank, Grade or Status of the Interviewer

For Army field test evaluations, it is recommended that the interviewer should be of similar rank or grade to the individuals being interviewed. A difference in rank or grade introduces a bias in the data which has been found to substantially influence test results. Interviewees tend to give the answer they perceive the higher-ranking interviewer favors. When the interviewer is of lower grade, the interviewee may not show respect and may not cooperate.

Evidence indicates that the greater the disparity between the status of the interviewer and that of the respondent, the greater the tendency for biased responses. Respondents tend to provide answers that will be more favorably received by the interviewer.

Data suggest that in the interview situation the respondent tends to support the norms adhered to by the interviewer. Lower socio-economic respondents may defer to the norms represented by a higher-status interviewer. The effect, however, is related to the types of questions asked. Sensitive issues involving socially accepted or rejected answers will effect more bias.

2. Sex of the Interviewer

Differences in response patterns according to the interviewer's sex depend on subject matter as well as on the composition of the respondent populations and other characteristics of the specific survey situation. Subject matter which tends to be most sensitive to differences in male/female response patterns deals with gender stereotypes. Interview items used in performance appraisals may be sensitive to sex role stereotypes. It is recommended that this type of item be investigated for rating differences between males and females. Interview items that are relevant to technical background experience (not usually obtained by females) also show gender response differences.

3. Race of the Interviewer

The effects of the race of the interviewer on the respondent should probably be viewed as the result of interaction between interviewer and respondent characteristics, or the result of the item content. Respondents often give socially-desirable answers to interviewers whose race differs from theirs, particularly if the interviewee's social status is lower than that of the interviewer and the topic of the question is threatening.

Nonsensitive, nonracial items appear to be relatively immune to interviewer effects for racial background. Therefore, racial background of the interviewer does not usually seem to affect survey results. It would be possible to assign interviewers of different racial background regardless of the respondent's racial background. An interviewer's race can probably establish different frames of reference for items with racially-related content. For threatening items or items with racially-related content, more valid results might be expected when the interviewer is of the same race as the respondent.

4. Experience of the Interviewer

There may be no significant differences between interview completion rates for experienced and inexperienced interviewers who have received sufficient interviewer training for face-to-face interviews and telephone interviews. However, it has been found that experienced interviewers may have different error rates than inexperienced interviewers. This error rate has been associated with the age of the interviewer, and the amount of interviewer training. Interviewer error is usually controlled through selection and training. Older interviewers (age 55 and over) have been known to frequently deviate from interviewer guides. Younger interviewers were found to follow the interview guides more closely. Nonstandardized administration of the interview could jeopardize the overall standardization of the survey procedures.

Other evidence indicates that field interviewers trained for less than a day produce more survey errors than more highly trained interviewers. Individuals responsible for developing interview items, guides, and training require sufficient development time prior to administration of the interview. Interview techniques to increase standardization have been known to improve through training. Response rates for telephone interviews may also be increased through training.

D. Situational Factors

Among the situational factors that should be considered when interviews are used are the following:

1. It helps greatly if the interviewees perceive the interviewer as interested in hearing their comments, as willing to listen, and (if the situation requires) as willing to protect them from recrimination for being adverse in their evaluations.

2. Interviews should be conducted in a quiet, temperature-controlled environment where the respondent can be comfortable and relaxed. Each respondent should be interviewed in private, separate and apart from all others, so that no other person hears or is biased by his/her responses.

3. The reinforcing behaviors of the interviewer have an influence on the responses collected, and at times may cause respondents to change their preferences. Such comments as "good" or "fine" and such actions as smiling and nodding can have a decided effect on test results. Praised respondents normally offer more answers than unpraised ones. Praising respondents may also tend to reduce "Don't know" answers without increasing insincere or dishonest responses.

4. Interested respondents seem to be more subject to interviewer effects than uninterested ones.

5. Interview questions which are read slowly indicate to respondents that they can take their time in carefully and thoughtfully answering the question. Rushing through an interview may reduce accuracy.

6. Use a "focus" group or pilot screening as a way to develop hypotheses and refine questions for establishing an interview guide and interview items. Interview guides are to be followed so that questions are asked without any wording changes. This promotes standardization across interviews.

7. Incomplete answers to survey questions require nondirective probing. When asking for clarification regarding an incomplete answer, the respondent is not to be directed toward any one response. Instead, phrases such as "tell me more" would be useful to employ.

8. Recording answers to interviews that use closed-end questions requires only that the interviewer mark the answer that the respondent selects.

9. When recording answers to open-ended questions, use a tape recorder if the respondent agrees or write down the answers verbatim. It is possible to combine open-ended and closed-end items for interview questionnaires, although coding and recording may be more difficult for the open-ended items.

10. For telephone surveys, use an interview structure and interview guide that promotes a high interaction between the interviewer and the respondent. This may be useful in increasing response rate.

11. Response cards can be adapted from face-to-face interviews for telephone surveys. Oral labeling of the scale points should be assessed on a pilot survey to be sure that the responses are not biased by the oral presentation of the scale.

E. Training Interviewers

Generally, interviewers require a certain amount of training. Army personnel may check with the Army Research Institute-Field Unit closest to them for help in this area. Some of the factors which should be considered when training interviewers are the following:

1. Training sessions for interviewers usually range between two days and five days. Interviewers conducting field interviews require more training than individuals who conduct telephone interviews. Two days minimum up through five days training are recommended for face-to-face interviews.

2. Sometimes researchers provide interviewers with information about the general research goals, sampling procedures, data analysis, and reports that will result from the survey.

3. Interviewer training requires general information in the course content such as how to introduce the study, as well as more specific information. Interviewers need to be familiar with the wording used in the survey, and any branching instructions. Standardization of the study through asking questions, probing incomplete answers, and recording answers are important aspects of the course content.

4. Interviewer training usually incorporates a demonstration of the standardized interview, and exercises where trainees role-play both the respondent and the interviewer. Practice sessions may also be tape recorded.

F. Data Recording and Reduction

In the structured interview, both questions and answers are orally communicated. The interviewer may encode the answers on paper, or tape record the responses for later encoding (but only if the interviewee agrees to the taping and does not seem influenced by the presence of a recording device).

Other topics related to interview data recording and reduction are outside the scope of this manual.

G. Special Interviewer Problems

This section notes some special problems related to interviews.

When interviews are used, the qualified interviewer will avoid leading, pressuring, or influencing the direction of an interviewee's evaluations. If potential interviewers have strong preferences regarding the system(s) being tested, they should probably be disqualified.

Many studies have been conducted that show other biasing effects on the interviewer. Factors leading to significant effects of the interviewer upon results include: relatively high ambiguity in the wording of the inquiry; interviewer "resistance" to a given question; and resistance to additional questioning or probing. Interviewer bias can exist without being apparent, and the direction of bias is not necessarily uniform. The least interviewer bias is probably found with questions that can be answered "Yes" or "No." The bias can result from differences in interviewing methods, differences in the degree of success in eliciting factual information, and differences in classifying the respondent's answers. Interviewers' expectations may have a more powerful effect on the results than their ideological preferences.

Some interviewers have a tendency not to transmit printed instructions word for word. Hence, total phrases may be eliminated and key words originally intended to focus the respondent's attention on some specific point are omitted or changed. Key ideas are lost, mainly through omission. Variability of interviewer performance seems to vary both across interviewers and within individuals.

An interviewer's attitude toward a question can communicate itself sufficiently to the respondent so that the meaning of the question is altered. When training interviewers to deliver a questionnaire in a standardized fashion, they need to rehearse the questions for tone of voice and body language to reduce any interviewer bias.